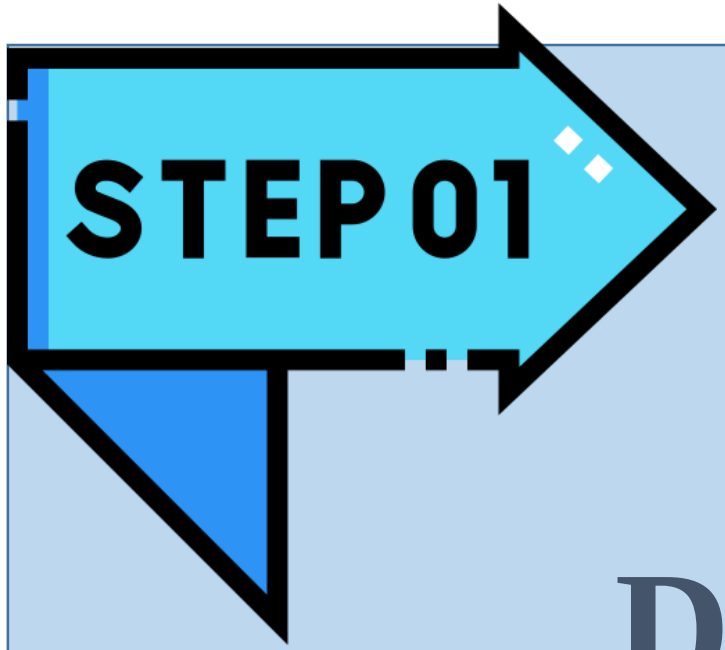
Data Science

Explanations about how to do the first and second phases of the final project of the data science course

University of Tehran
Spring 2024

Mohammad Amanlou
Amir Mahdi Farzaneh
Shahzad Momayez

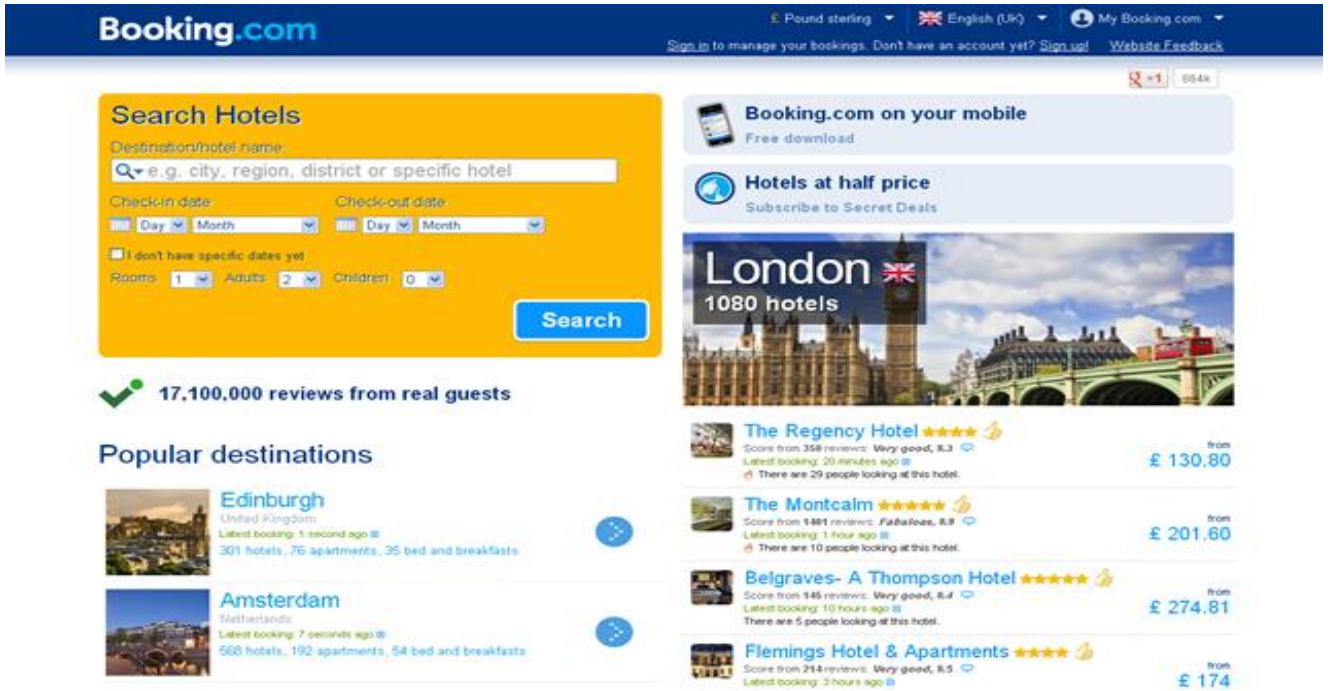


Data Scraping

Data Scraping

We used www.booking.com

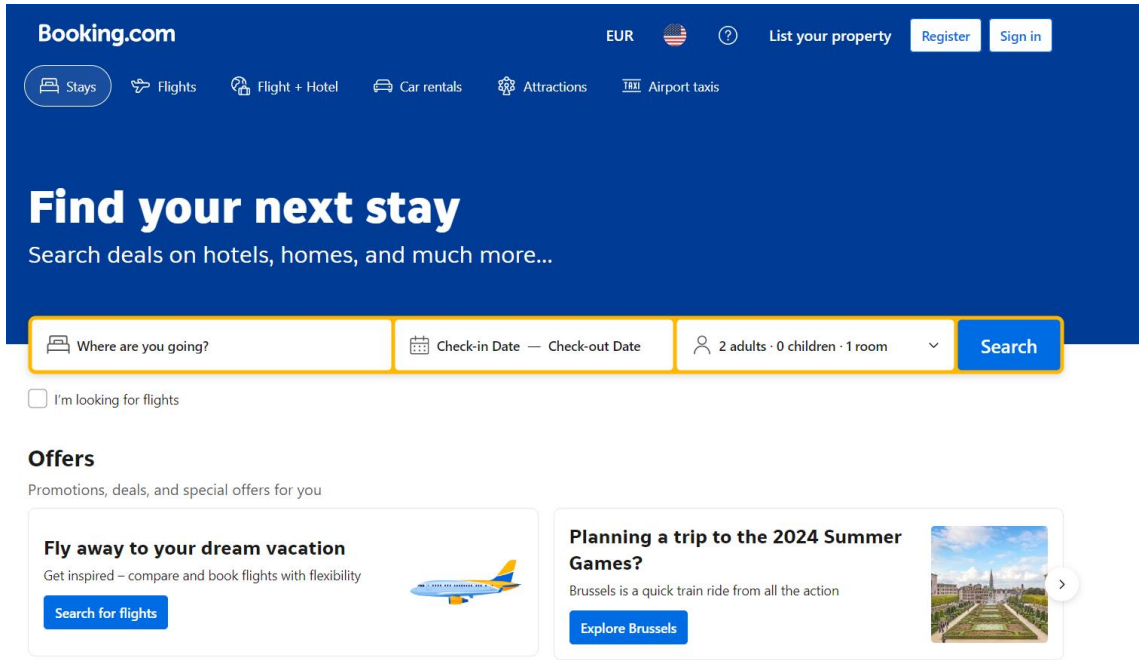
Contains reservation information of hotels

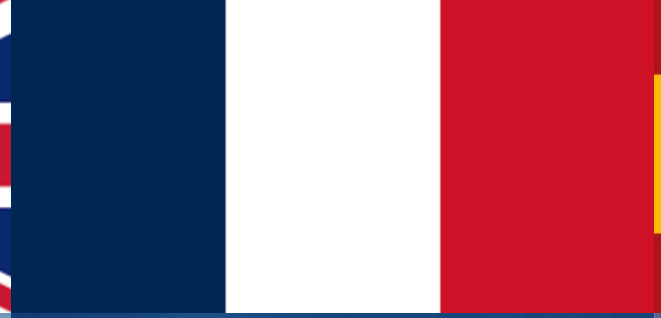


Data Scraping

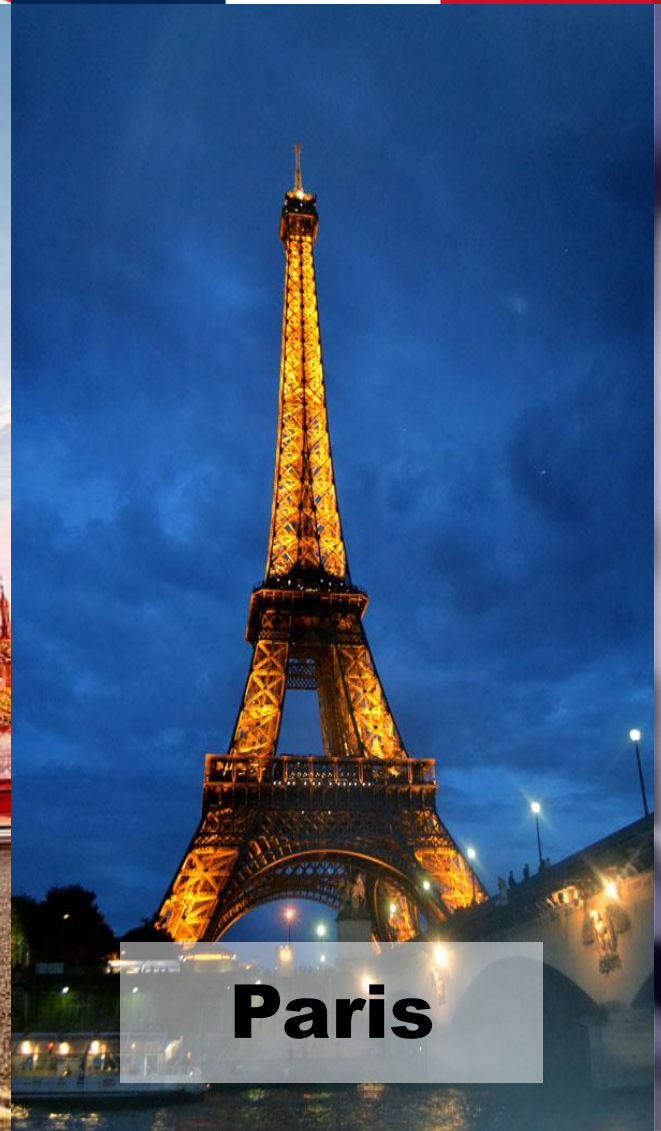
We used beautiful soup library for scrapping

Our dataset contains about 8000 samples





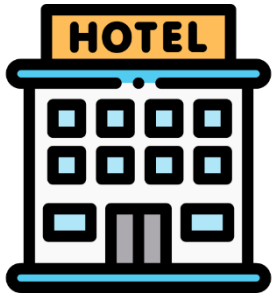

London


Paris


Madrid


Berlin

Data Scraping



It is the process of extracting data especially from websites. It involves using automated tools or scripts to collect large amounts of data.

- **Scrape Hotels**:
- Initialize an empty list all_hotels to store hotel data.
- Enter a loop to **handle pagination**:
  - Send a **GET request** to Booking.com with the defined headers and parameters.
  - Parse the HTML response with **BeautifulSoup**.
  - Extract hotel data from the current page using get_hotels_from_page.
  - Add extracted hotel data to all_hotels.
  - Increment the offset parameter to move to the next page.
  - Break the loop if no more hotels are found or after 20 pages.

Data Sample

hotel_name	location	price	room_type	beds	rating	rating_title	number_of_ratings	per night	Log_number_of_ratings	Log_price	Size
Aloft Riyadh Hotel	Riyadh	475	Breezy Room, Guest room, 1 King	1	8.2	Good	2947	1	7.988543	6.163315	32
Nourth House ApartHotel	Al Qurayyat	200	Budget Twin Room	2	7.5	Good	264	1	5.575949	5.298317	35
Dyar Al Hamra Hotel	Jeddah	340	Junior Suite King	1	7.7	Good	3923	1	8.274612	5.828946	28
Courtyard by Marriott Riyadh Northern Ring Road	Riyadh	525	Standard, Guest room, 1 King	1	8.1	Good	1032	1	6.939254	6.263398	30
Golden Bujari Al-Dhahran - Hotel	Al Khobar	490	Deluxe Twin Room	2	8.4	Good	2315	1	7.747165	6.194405	40

Description of all columns of the df

- name: shows the name of the hotels in europe
- location: indicates the "city" and the "neighbourhood" of that hotel
- price: shows the price of hotel in US$.
- rating: indicates the rate of the hotel from 10.
- quality: shows the quality of the hotel
- review: shows how many times this hotel has been reviewd.
- bed: shows the number of the beds it has.
- distance from centure: shows the distance of the from centure of the city.
- room_type: shows the type of the room in the hotel that had been reserved.
- nights: indicates the number of nights the room has been reserved.
- adults: shows the room has been reserved for how many adults.

Feature Engineering

Feature Engineering

- We focused on extracting specific components from "location" coulumn that containd 2 different information about the neighborhood and the city.
- so we replace "location" with 2 different columns:
    - neighborhood
    - city

Feature Encoding

- We numbered and encoded the categorical columns of our data, which had descriptive values equal to the level of customer satisfaction, according to the average rating available for each descriptive category and the concepts we know from the English language.

```
mapping = {'Review score' : 0, "Guest rating" : 0 ,'Good' : 1, 'Very good' : 2,'Fabulous' : 3, 'Superb' : 4 , 'Exceptional' : 5}
df['quality_val'] = df['quality'].map(mapping)
```

- Then we converted the values that were stored in other columns dirty with extra punctuation marks or as strings.

Null Handling

- We examined several methods to fill null values, among which the following methods can be mentioned:
  1. Sorting based on address and using forward fill or back fill, this method did not create suitable values for us
  2. Using imputers such as KNN or RF to fill null values, these methods also did not provide good accuracy due to the high overhead of filling null values, as well as the number and distribution of null values in the data, so they were not used.
  3. Filling the missing values with mean, median and mode, this method was used as a safe traditional method in the first step.
  4. The use of interpolation methods, which also provided us with good accuracy.
  5. Using the random sampling method from the column with zero values, this method also did not give us good accuracy because the zero values in the categorical columns were few.

Null Handling

Initial df

name	0
location	0
price	0
rating	193
quality	153
review	153
bed	23
distance from centure	0
room_type	0
nights	0
adults	0
city	0
neighbourhood	0
quality_val	0

final df

name	0
rating	0
quality	0
room_type	0
city	0
neighbourhood	0
quality_val	0
nights	0
adults	0
review	0
price	0
bed	0
dfc	0

dtype: int64

Explanatory Data Analysis

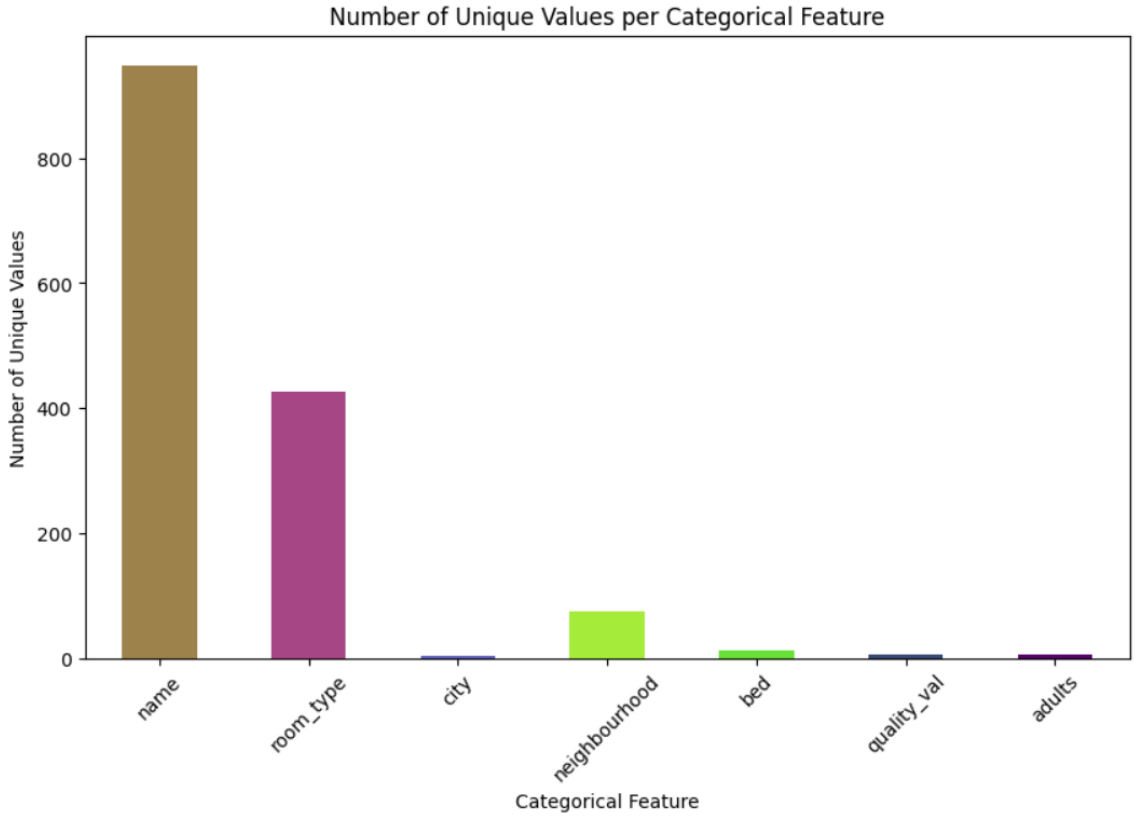
EDA

- To find the attributes of each city and the attributes of the medita set, we grouped by city, point, and city and calculated different attributes for each attribute in that city.

	price				
	mean	max	min	std	median
city					
Berlin	372.451596	2589	15	337.246117	279.5
London	456.130271	11507	26	623.650952	283.0
Madrid	817.373500	7305	10	979.460679	570.0
Paris	459.115261	4289	21	464.511245	328.0

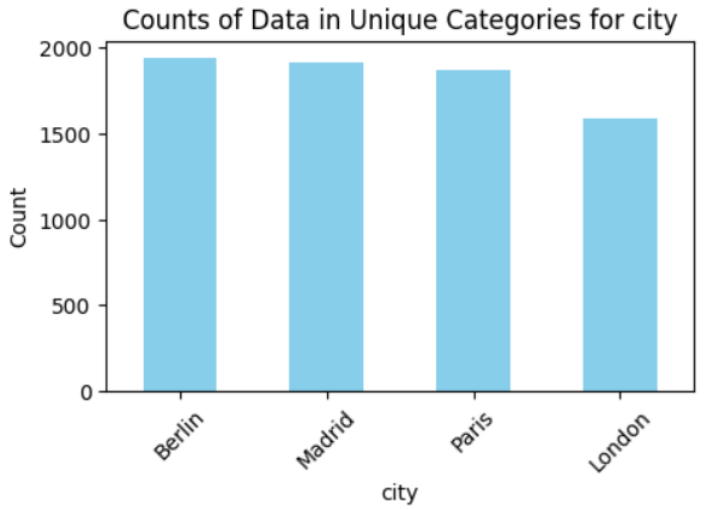
EDA

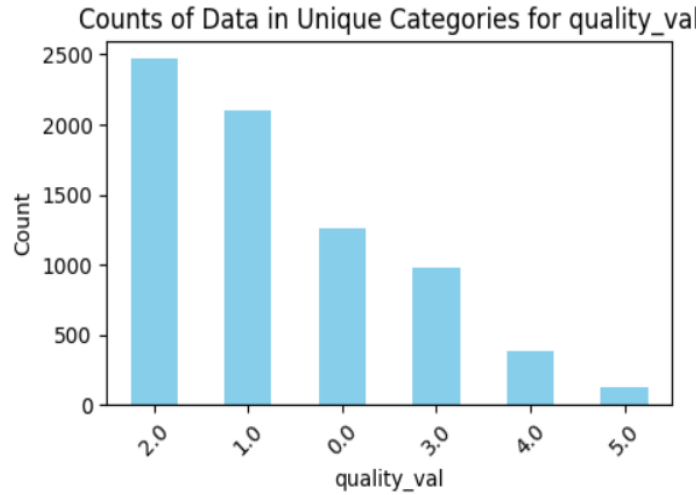
- In this section, we calculated the number of unique values in each column in order to know if the number of these values is low or high, and also to identify the categorical ones.

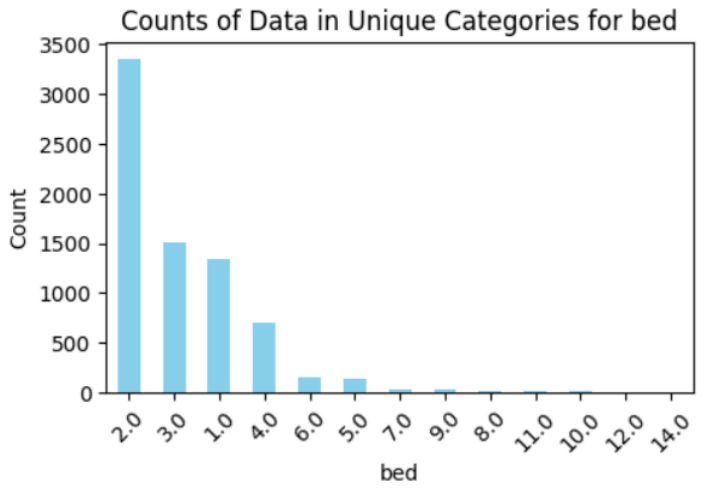


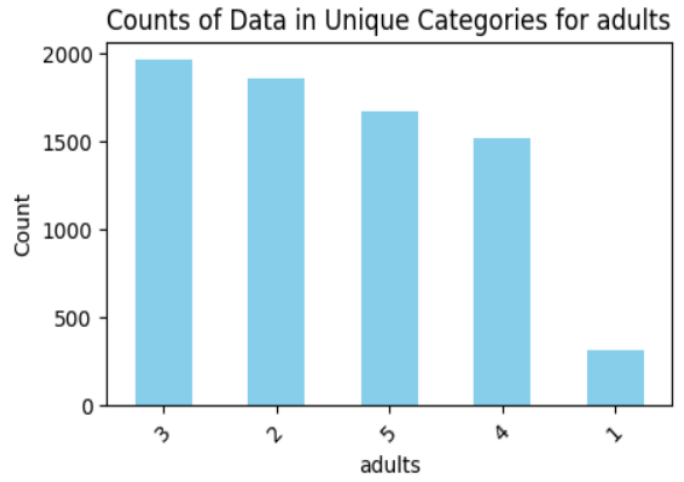
EDA

- Now, for those columns that had a limited number of categories, we also calculated the number of data in each category.



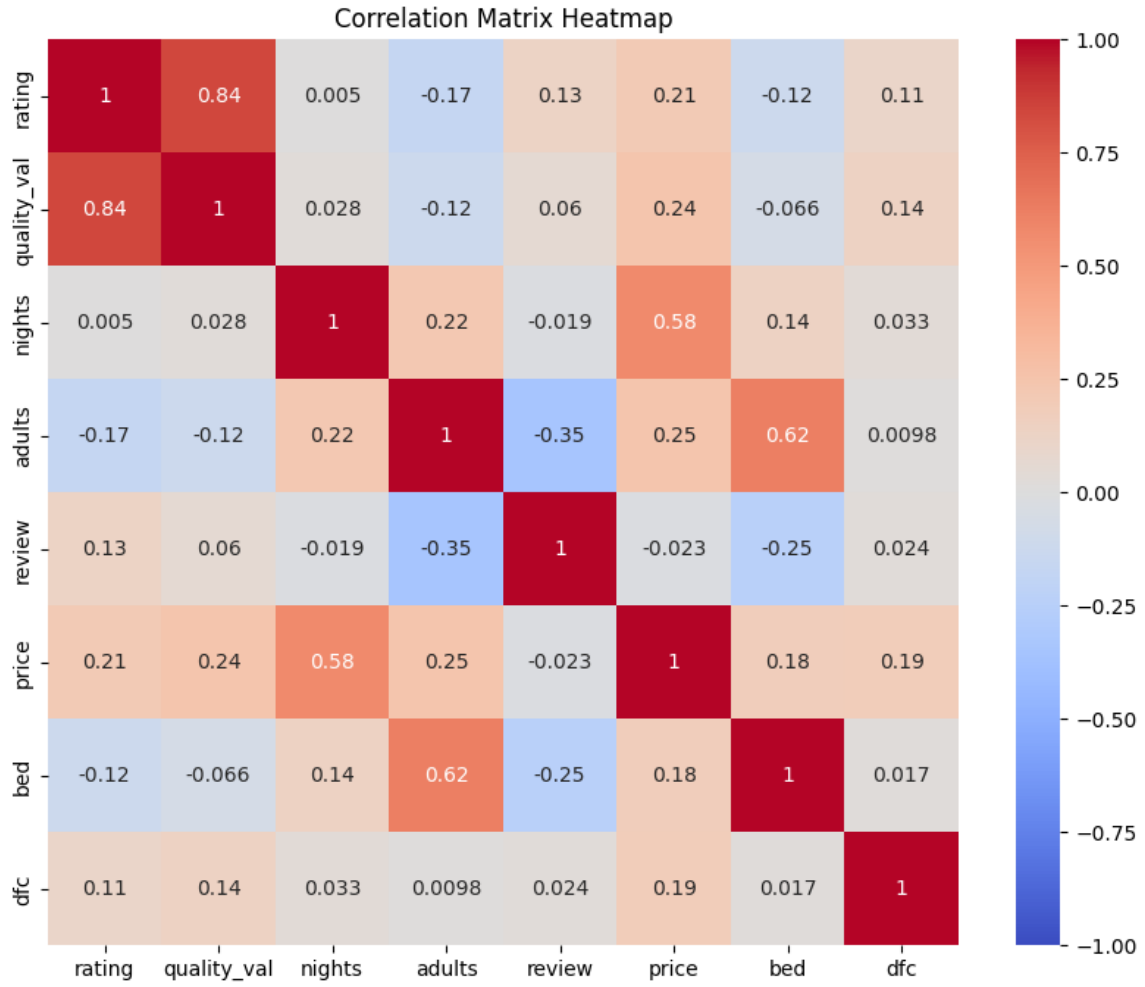






Correlation Matrix

- The heatmap provides a quick visual summary of the relationships between all variables, allowing you to easily identify strong positive, strong negative, or weak correlations.
- Identify patterns: You can see patterns in the correlations, such as groups of variables that are highly correlated with each other.



Monte Carlo Simulation

- **Price Analysis:** Monte Carlo simulation can help generate different possible price scenarios based on the variability in the data. This can help in predicting the range of prices for different combinations of variables.

```
num_simulations = 1000
simulated_prices = []
for _ in range(num_simulations):
    # Generating random samples for analysis
    simulated_prices.append(np.random.normal(df['price'].mean(), df['price'].std()))

simulated_prices_mean = np.mean(simulated_prices)
simulated_prices_std = np.std(simulated_prices)

print(f'Mean simulated price: {simulated_prices_mean}')
print(f'Standard deviation of simulated price: {simulated_prices_std}')
```

```
Mean simulated price: 559.3415418237078
Standard deviation of simulated price: 674.098866361682
```

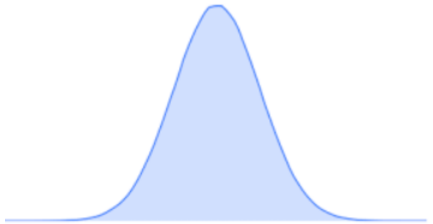
Monte Carlo Simulation

- **Quality Analysis:** By simulating the quality variable in the DataFrame, Monte Carlo simulation can provide insights into the distribution of quality ratings, helping to identify patterns or outliers.

```
num_simulations = 1000

# Create empty lists to store simulation results
simulated_quality = []
for _ in range(num_simulations):
    # Generating random samples for analysis
    simulated_quality.append(np.random.choice(df['quality']))

# Analyze the simulated quality data
quality_counts = pd.Series(simulated_quality).value_counts()
quality_percentage = quality_counts / quality_counts.sum() * 100

print("Simulated Quality Distribution:")
print(quality_percentage)
```

```
Simulated Quality Distribution:
Very good       32.6
Good            30.5
Review score    17.4
Fabulous        12.4
Superb           5.2
Exceptional      1.8
Fabulous 8.9     0.1
```

Central Limit Theorem



- **Rating**: 95% Confidence Interval: (8.275, 8.363)
  - **Inference**: We can be 95% confident that the true average rating of listings in the population is between 8.275 and 8.363. This indicates that the average rating is quite high, around 8.3.
- **Quality_val**: 95% Confidence Interval: (2.157, 2.295)
  - **Inference**: The true average quality value is likely between 2.157 and 2.295. This suggests a moderate quality rating on the given scale.
- **Nights**: 95% Confidence Interval: (1.430, 1.510)
  - **Inference**: The average number of nights for a booking is between 1.430 and 1.510 nights. This suggests that most bookings are short-term, typically around 1.5 nights.
- **Adults**: 95% Confidence Interval: (1.472, 1.552)
  - **Inference**: The true average number of adults per booking is between 1.472 and 1.552. This indicates that bookings are usually made for about 1.5 adults on average, often just one or two people.
- **Review**: 95% Confidence Interval: (4408.453, 5254.196)
  - **Inference**: The true average number of reviews per listing is between 4408.453 and 5254.196. This high number suggests that listings are well-reviewed, indicating potentially high popularity or many transactions.
- **Price**: 95% Confidence Interval: (284.153, 326.318)
  - **Inference**: The true average price of a listing is between 284.153and284.153and326.318. This range gives us a good estimate of the typical cost, which appears to be in the mid-300 range.
- **Bed**: 95% Confidence Interval: (1.102, 1.205)
  - **Inference**: The average number of beds per listing is between 1.102 and 1.205. This suggests that most listings offer slightly more than one bed, typically suitable for small groups or solo travelers.
- **Distance from Center (dfc)**: 95% Confidence Interval: (2.555, 8.570)
  - **Inference**: The average distance of listings from the city center is between 2.555 and 8.570 kilometers. This wide range indicates that listings vary significantly in their proximity to the center, with some being quite central and others farther away.

Hypothesis Testing

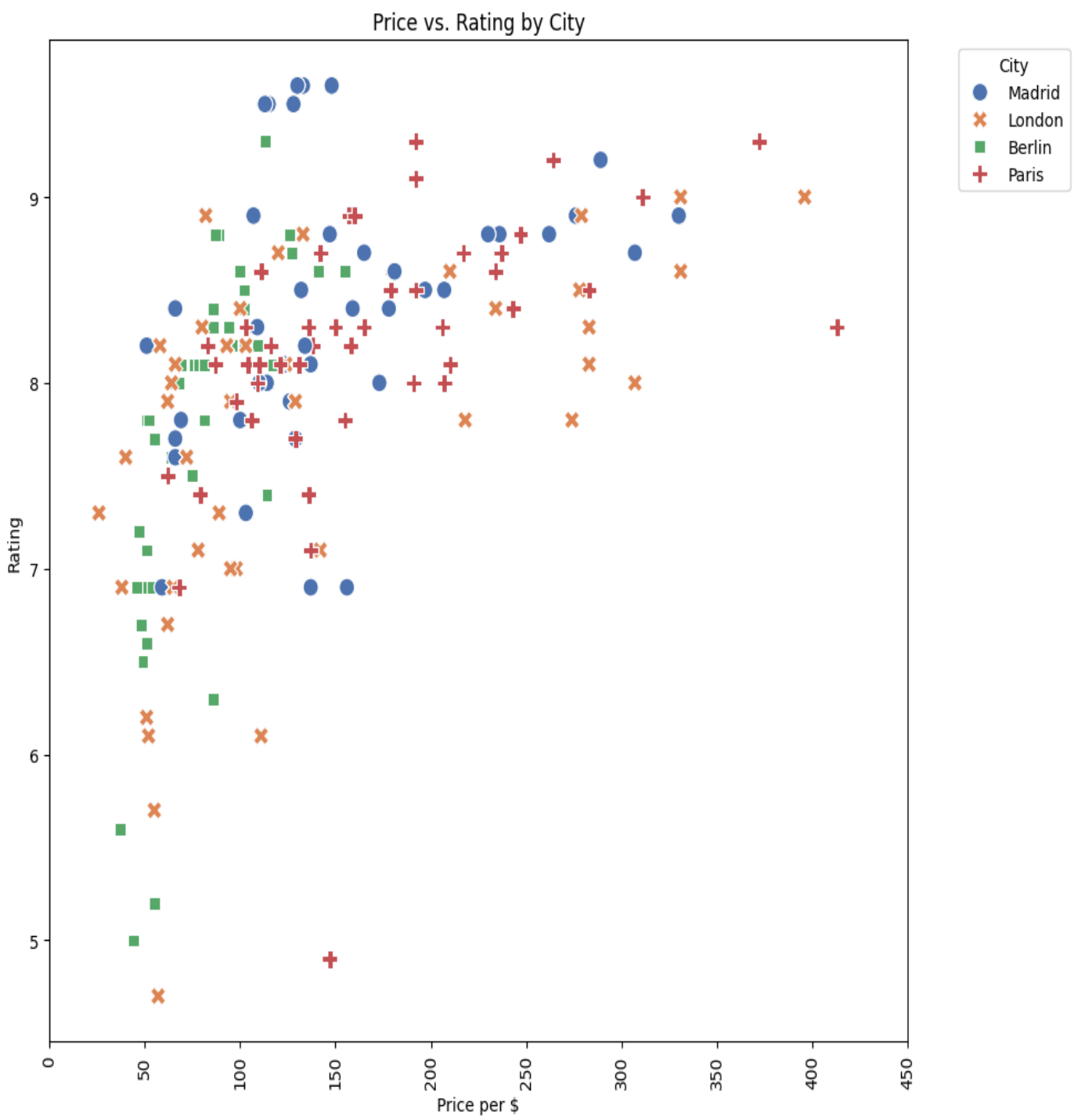
- In this part our null hypothesis was distance from center of city has impact on price and rating.
- P-value for both is less than 0.05
- So null hypothesis rejected and we have linear relation between center of city and price also rating.

Visualization

Price and rating
For each hotels' cities

Madrid has better hotels

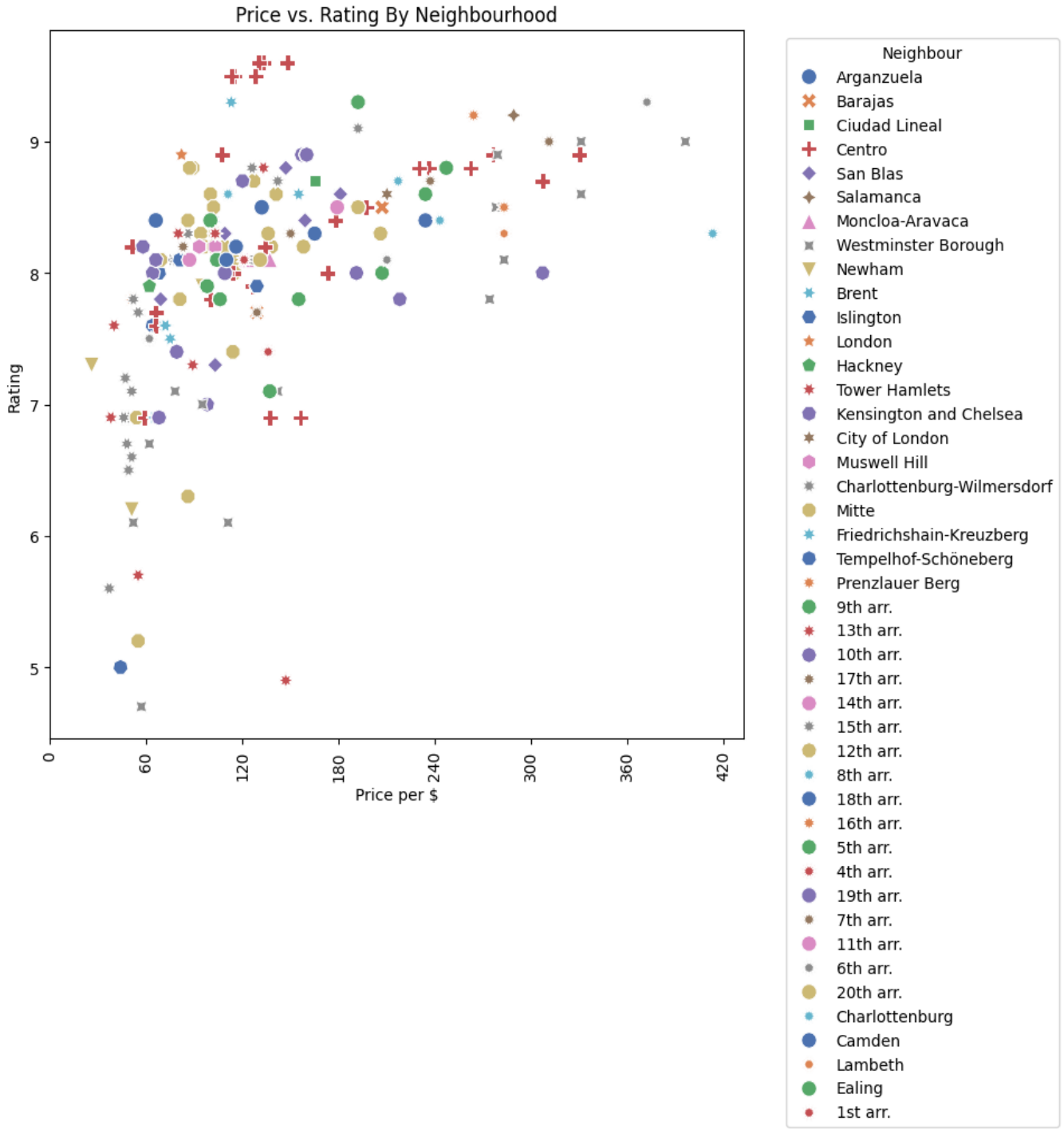
London ,The most expensive hotels



Visualization

All neighbors

We suggest Centro of Madrid Westminster brough ,The most expensive hotels



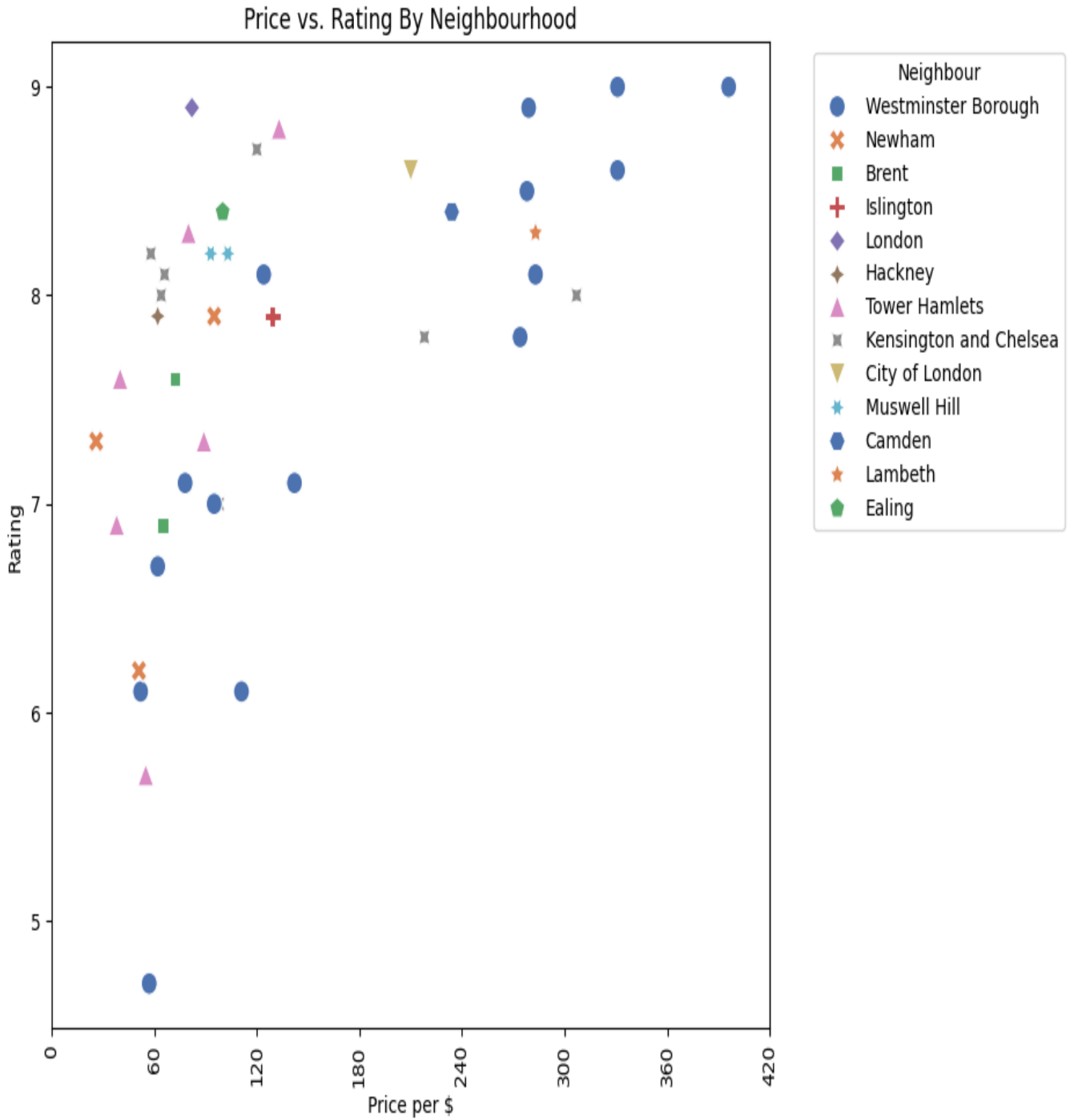
Visualization

London

Westminster brough ,The most expensive neighbors

Also low rating and cheap hotels are here

Other neighbors are appropriate



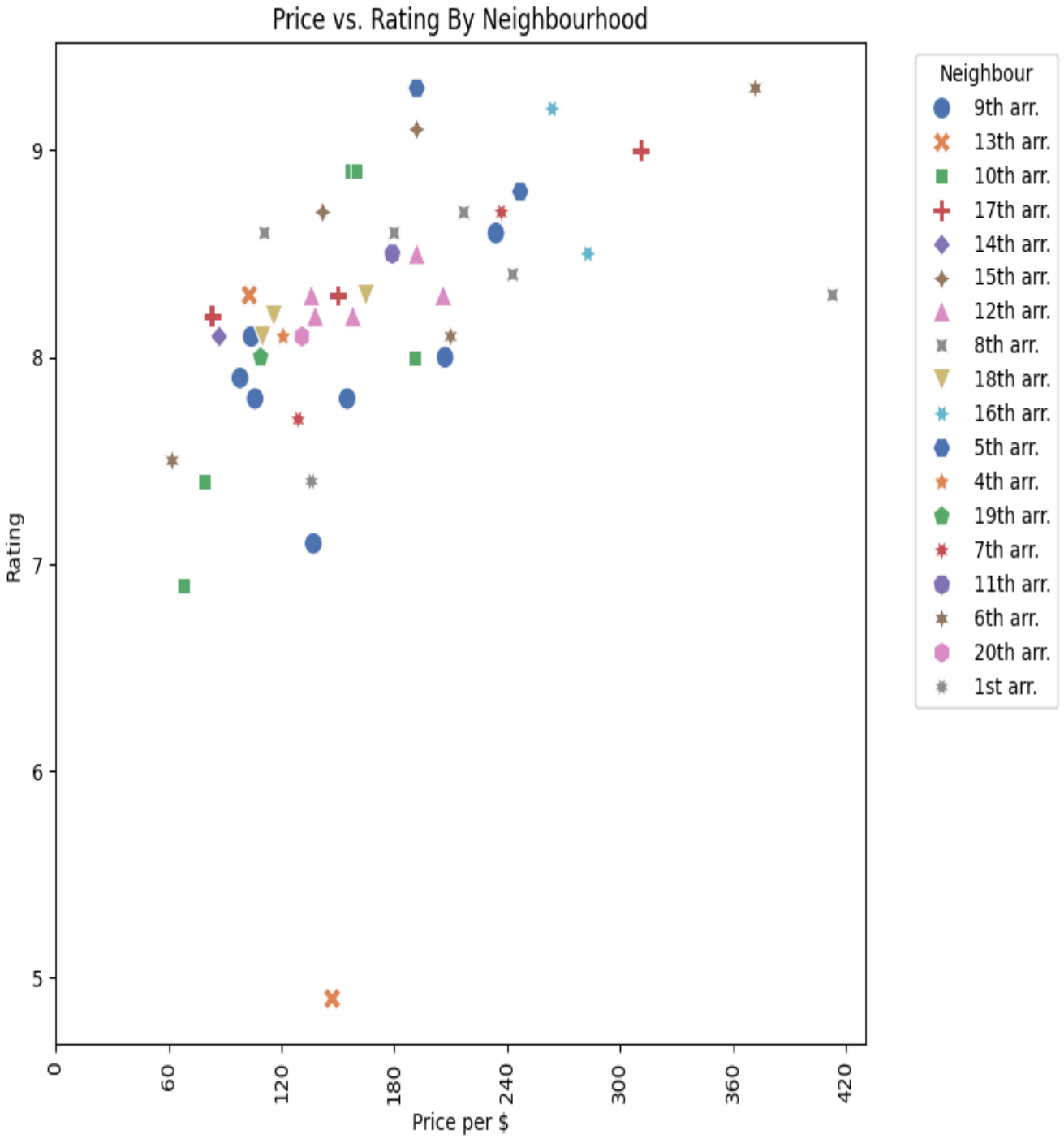
Visualization

Paris

The most expensive is 16h arr.

The cheapest is 10th arr.

All neighbors sound good for booking

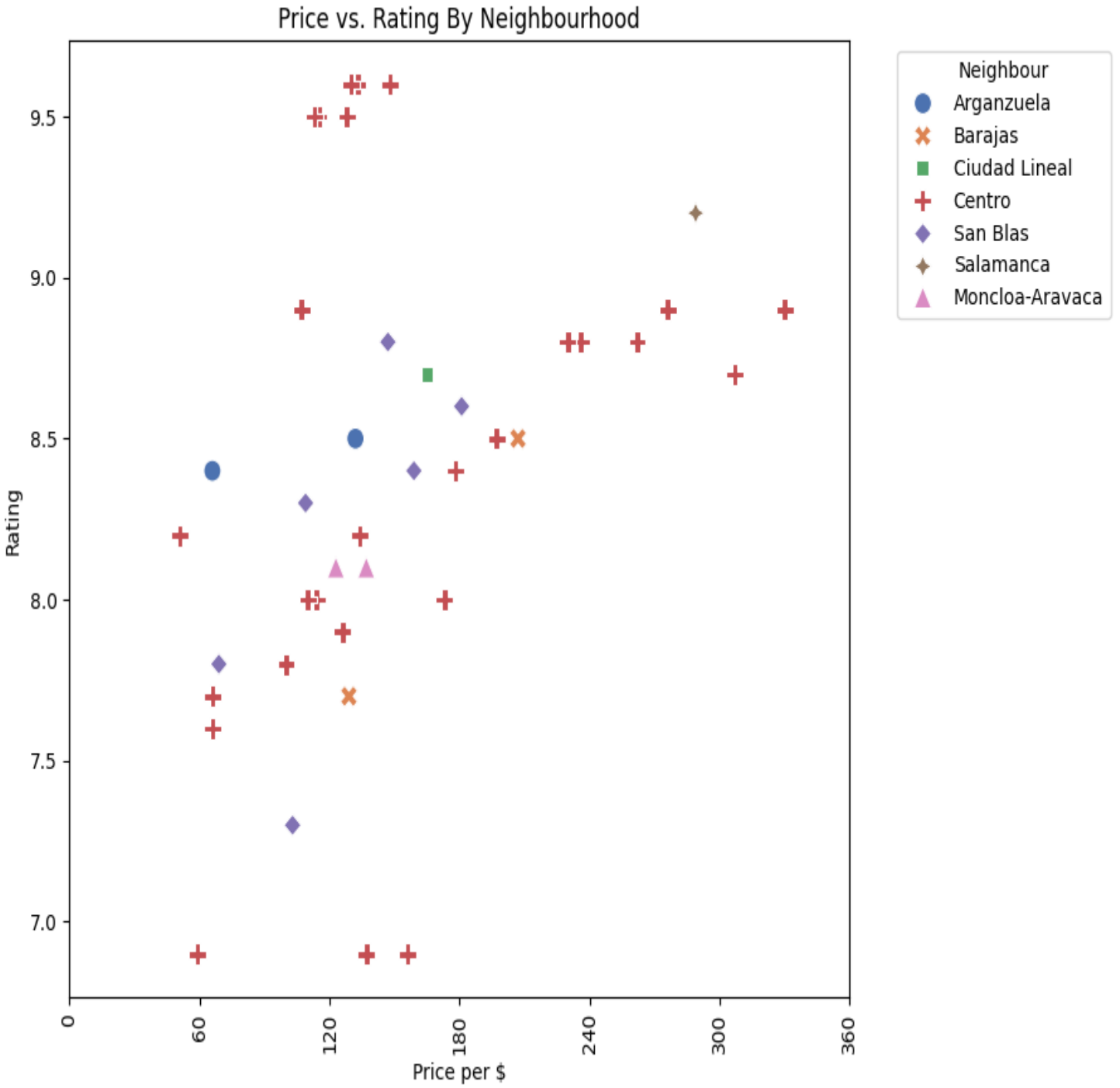


Visualization

Madrid

Most of information are fro centro

Also the best hotels are ove there



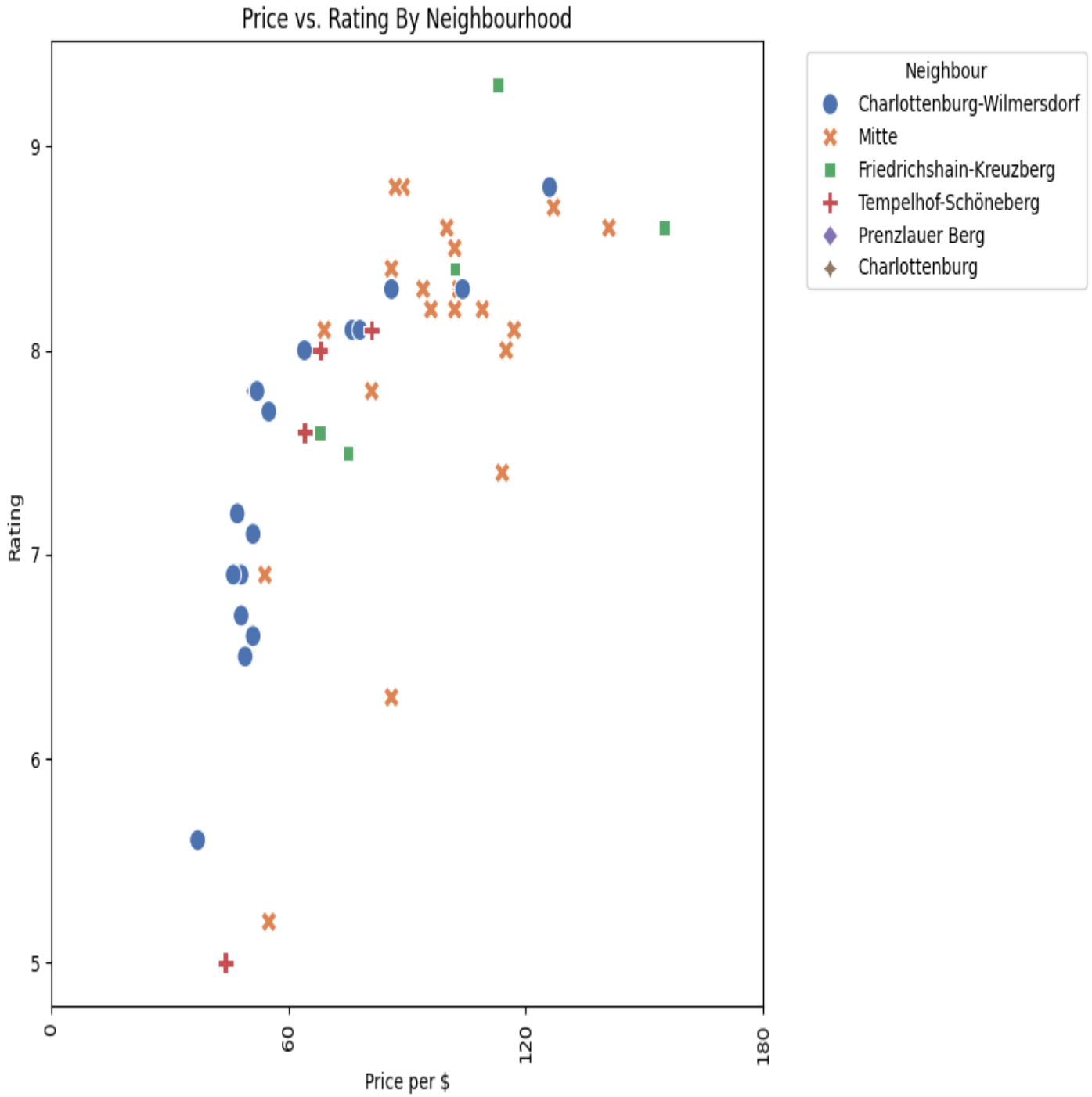
Visualization

Berlin

Mitte has better hotels

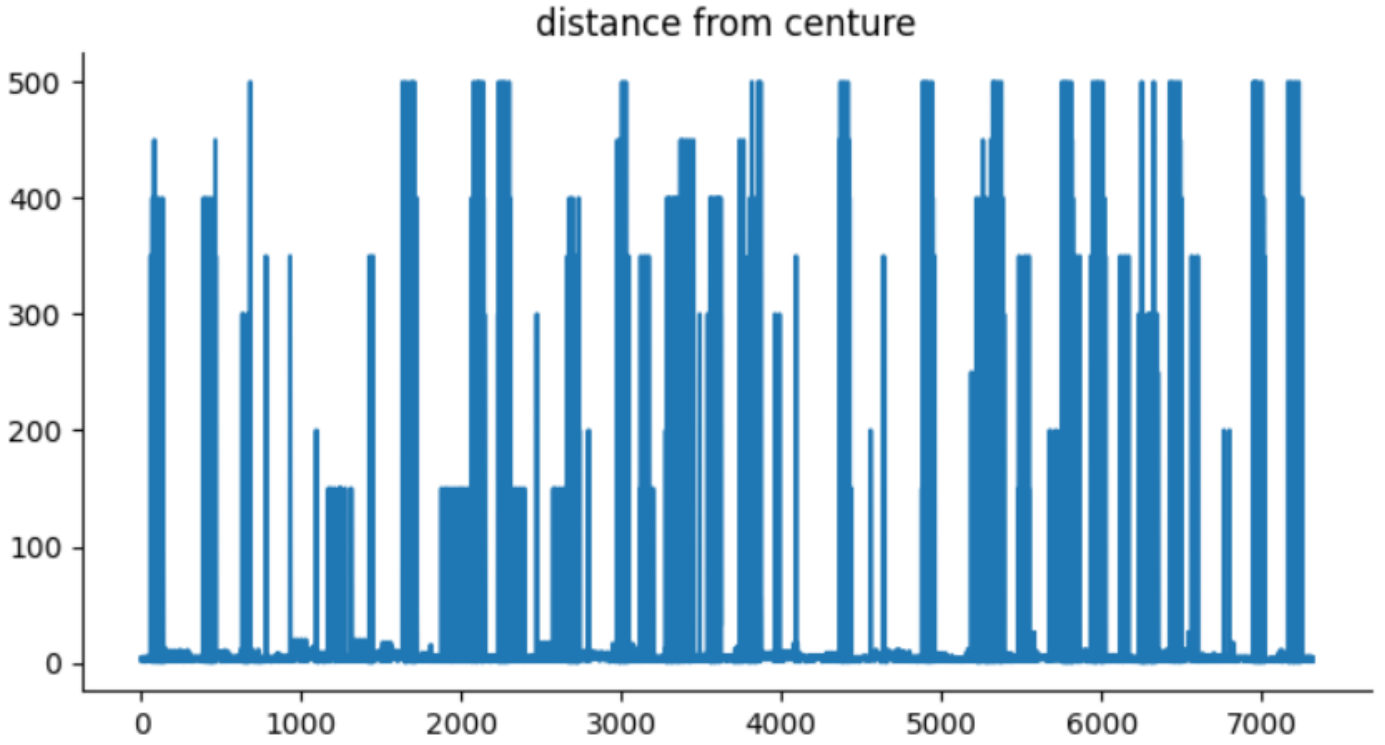
Also

Charlottenburg is not good idea



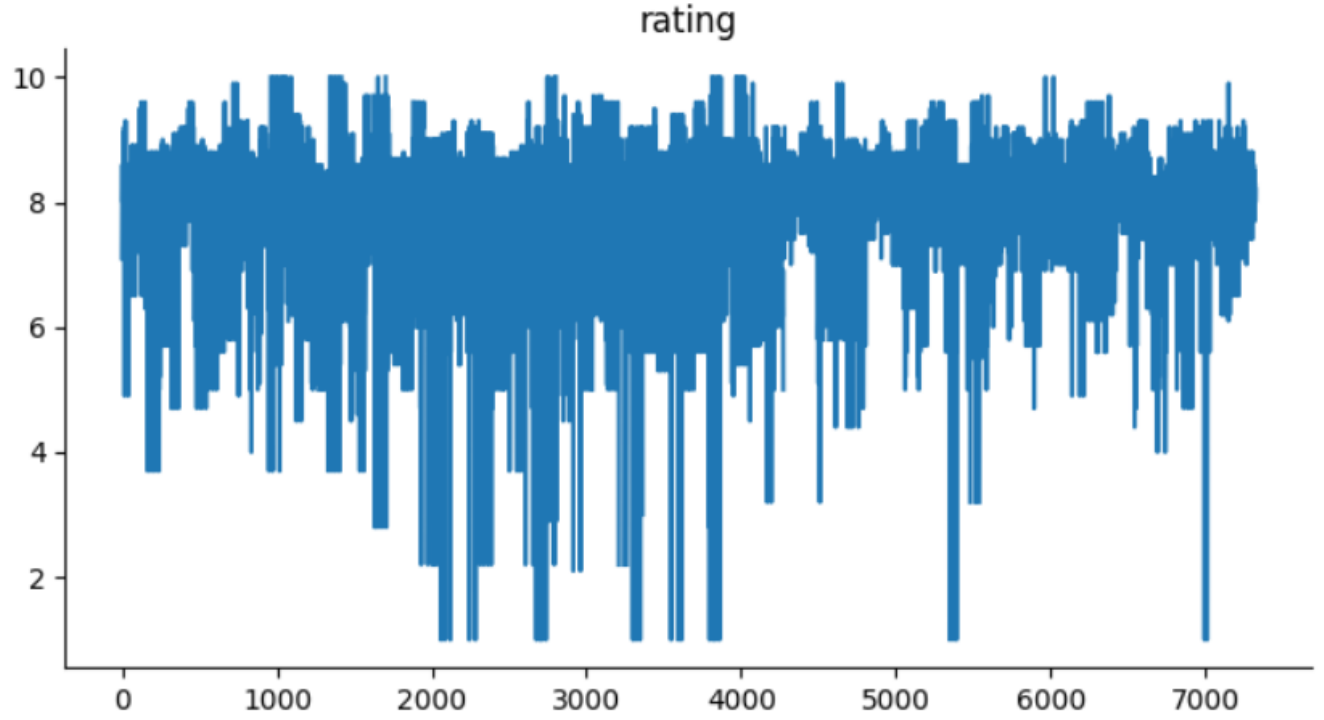
Visualization

- it is a line graph titled "distance from center". The x-axis is labeled "each data" and the y-axis is labeled "counts".
- The line graph displays the counts of data points at various distances from the center. The distance starts at zero and goes up to 600 in increments of 100.
- However, we can see some general trends. The counts appear to be highest at a distance of 100 and then steadily decrease as the distance from the center increases. There appears to be a bit of a bump in counts around 400 distance units from the center. most of the hotels are located in the center of the city so that the distance from the hotels to the center of the city in most cases is equal to 0.



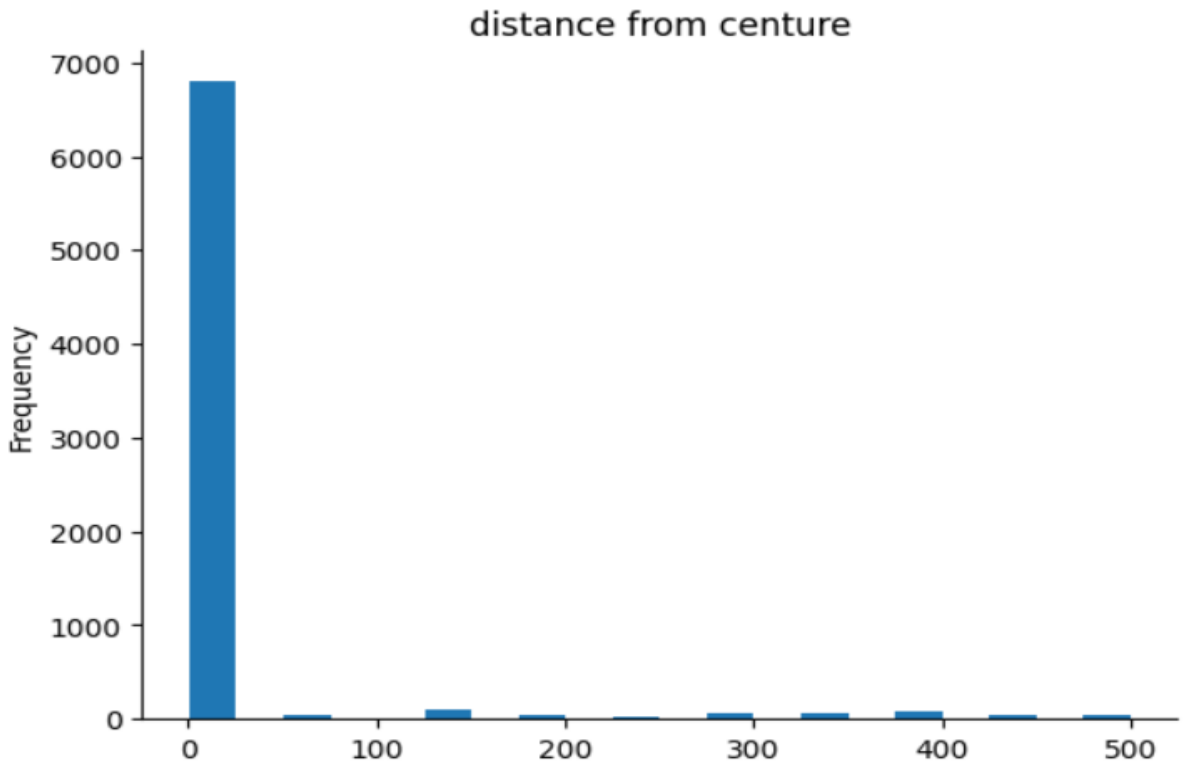
Visualization

- In the specific image, the x-axis appears to represent each data and the y-axis appears to represent the value.
- this show that the rating of the hotels are mostly between 7 - 9 but there are some hotels that are out of range and can be considered as outliers. for example one of tham has rating less than 5.



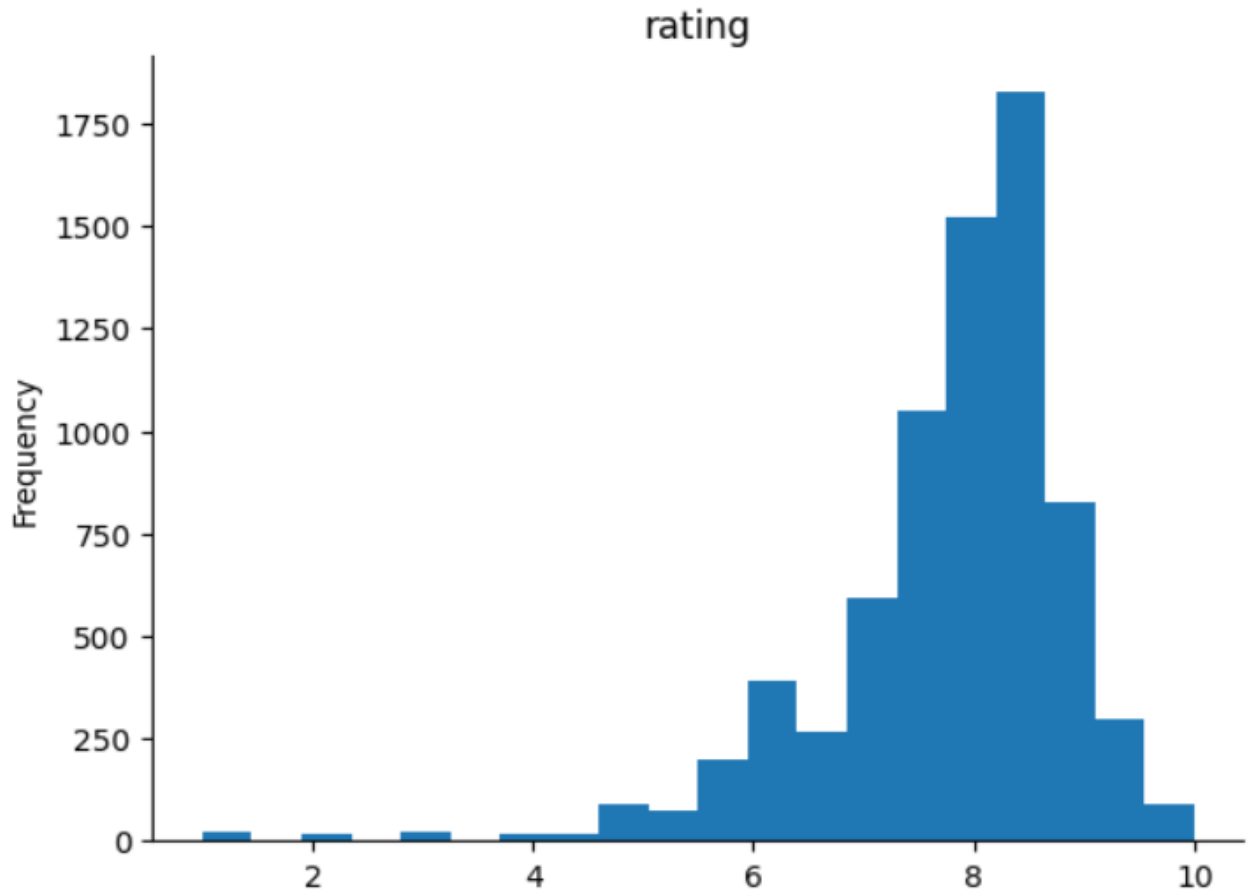
Visualization

- it is a bar chart titled "distance from center". The x-axis is labeled "the hotel data" and the y-axis is labeled "distance from center". There are data points plotted for every week from 0 to 40.
- The barchart shows the average distance from the center of a city. it shows that most of the hotels have 0 distance from the center of the city. in other words, most of them are located in the city center.



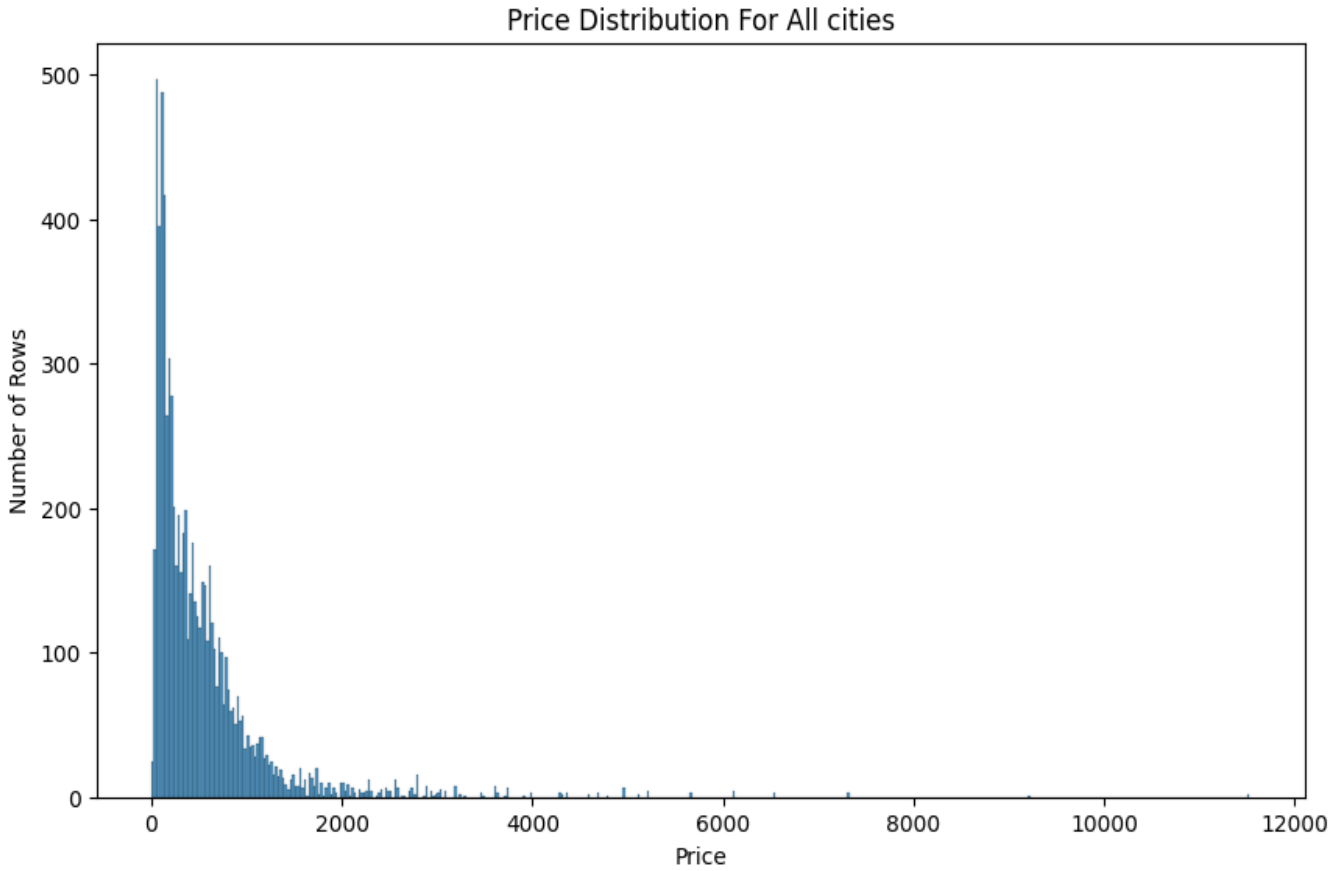
Visualization

- this chart shows that the rating of the hotels has mostly been between 7 to 9. and are mostly between these numbers. here we can see some outliers that in some cases rating of a hotel has been 5 or some other number. and also in some points it has the most rating. the frequency of rating between 8-9 has been the most.



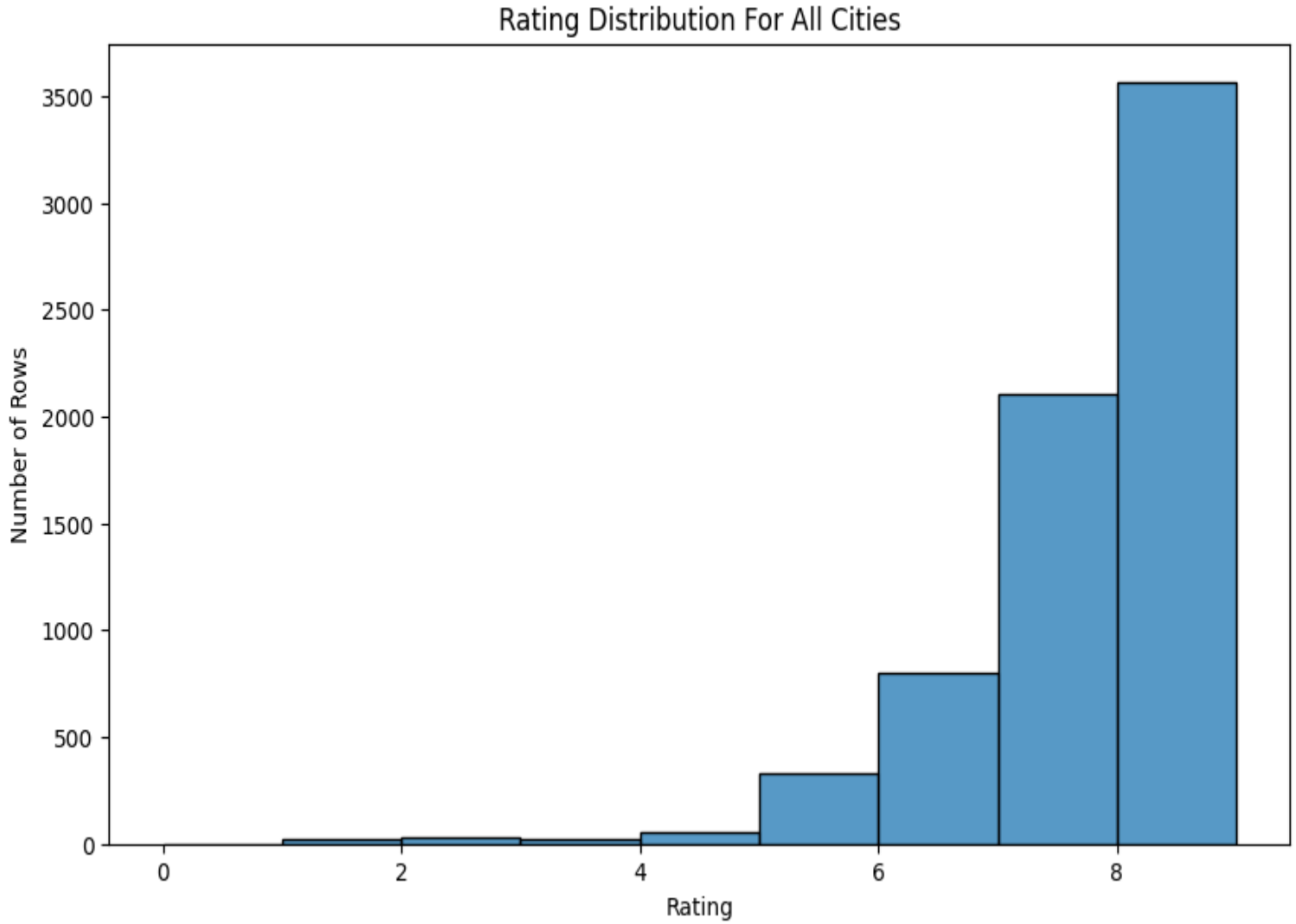
Visualization

- It is histogram plot for price all hotels in these four cities.
- As we see it is right skewed like salary distribution.



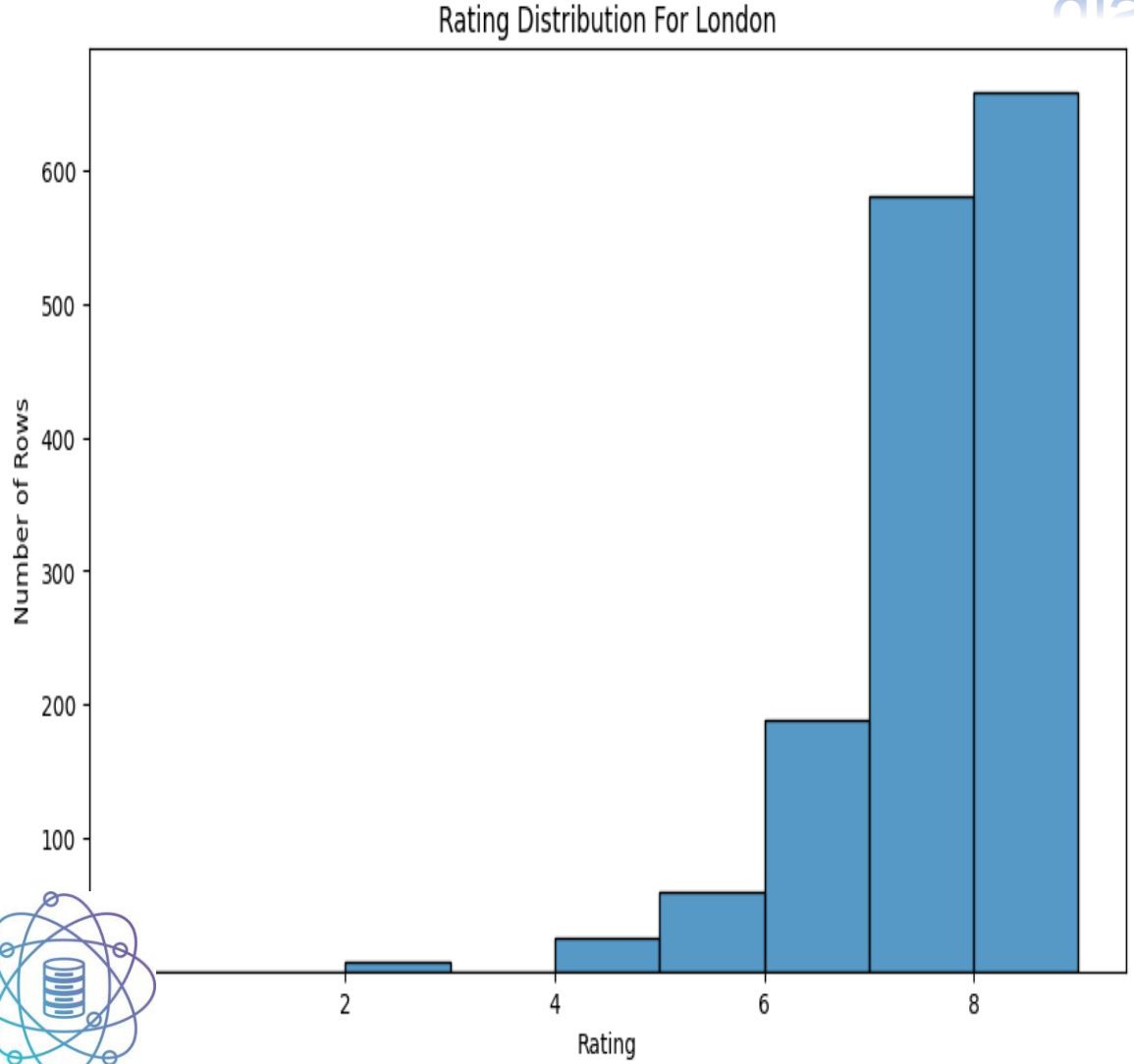
Visualization

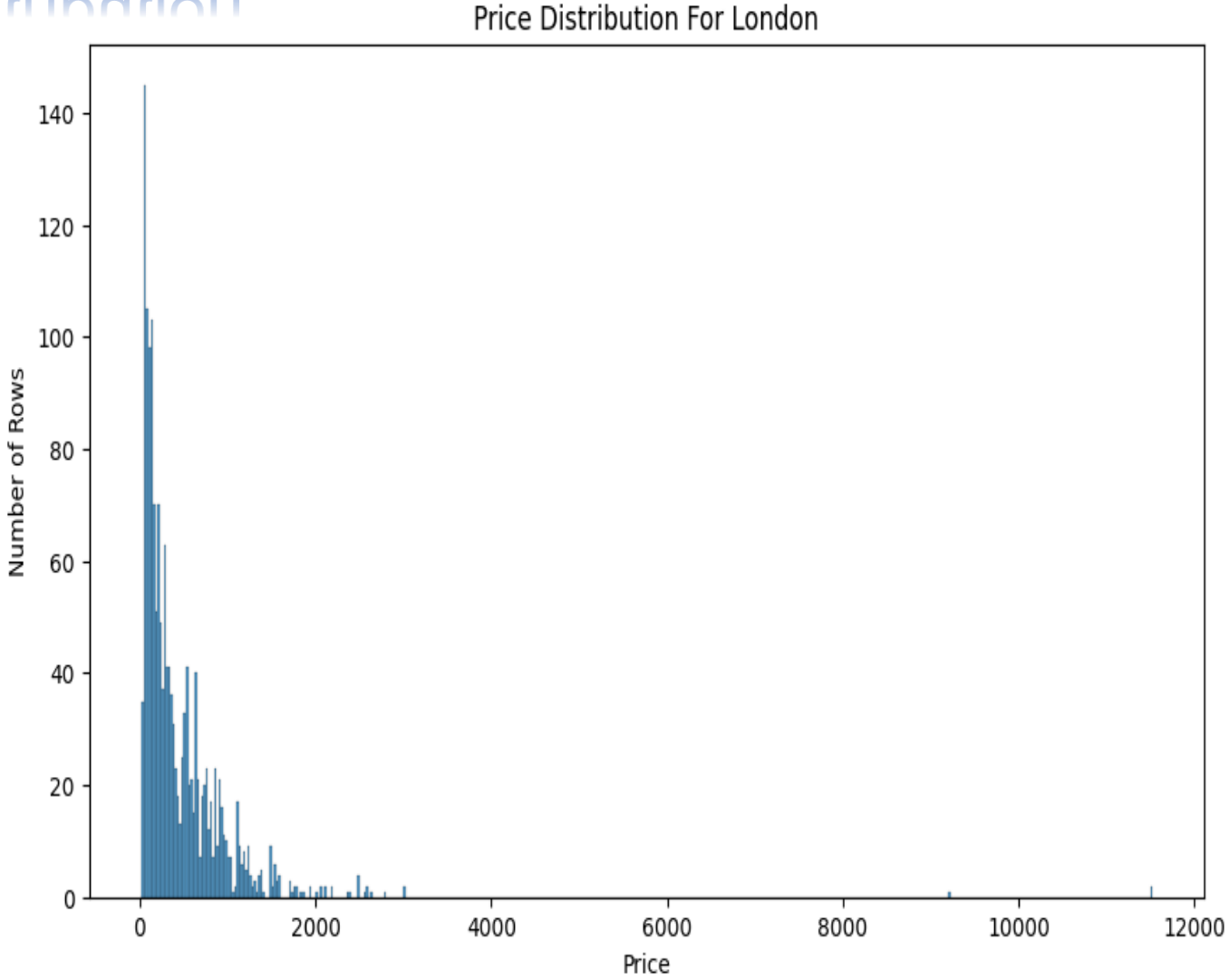
- It is histogram plot for rating all hotels in these four cities.
- As we see it is left skewed



London

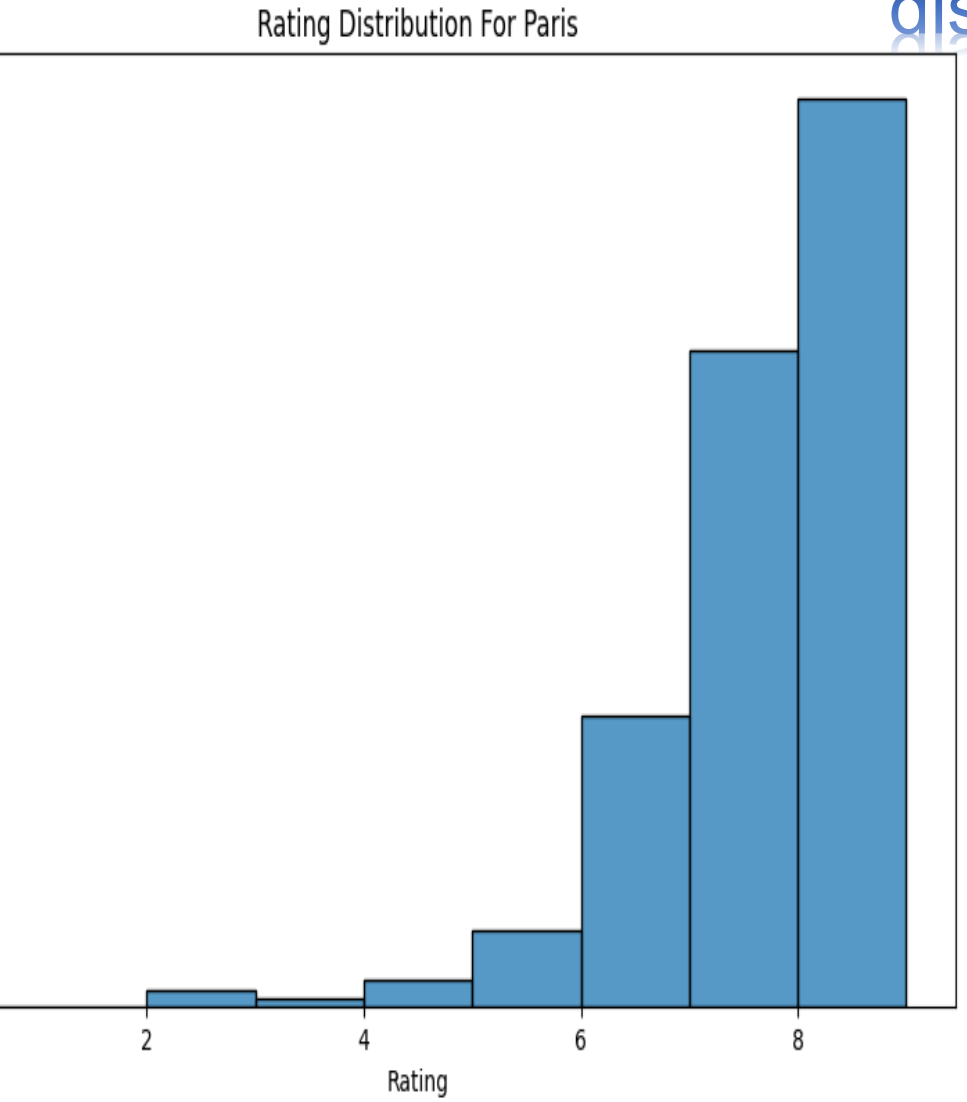
Rating and Price distribution

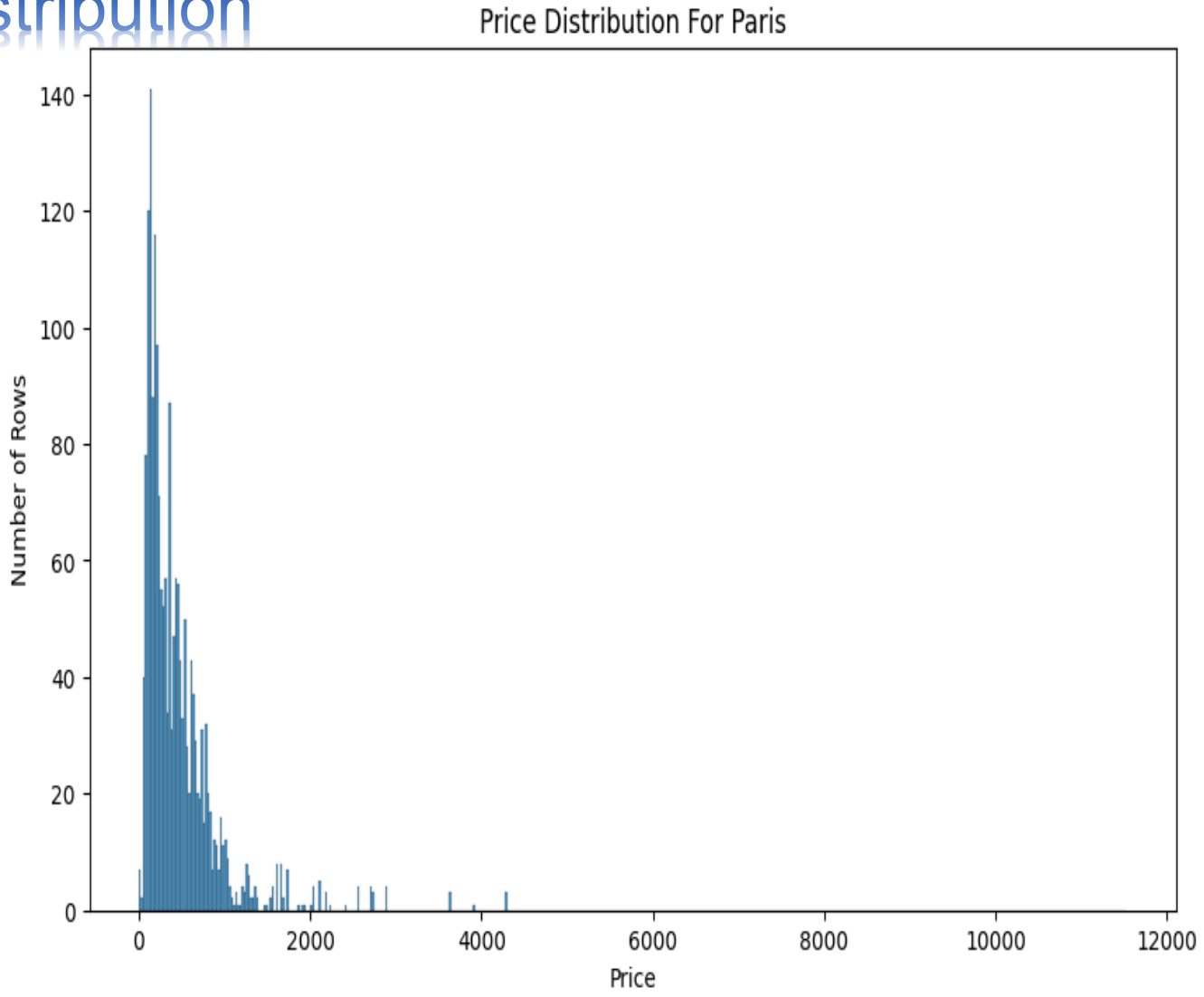




Paris

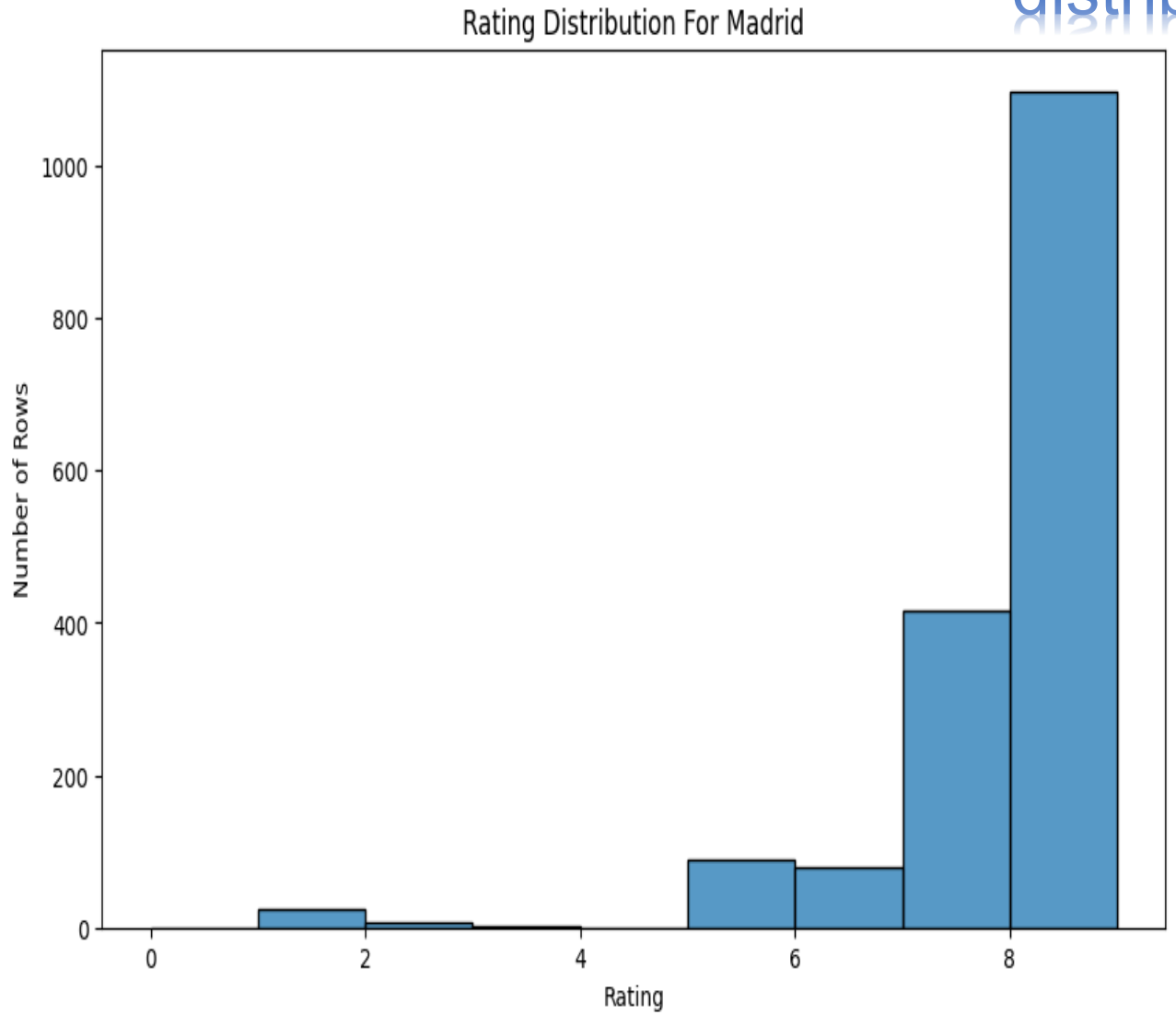
Rating and Price distribution

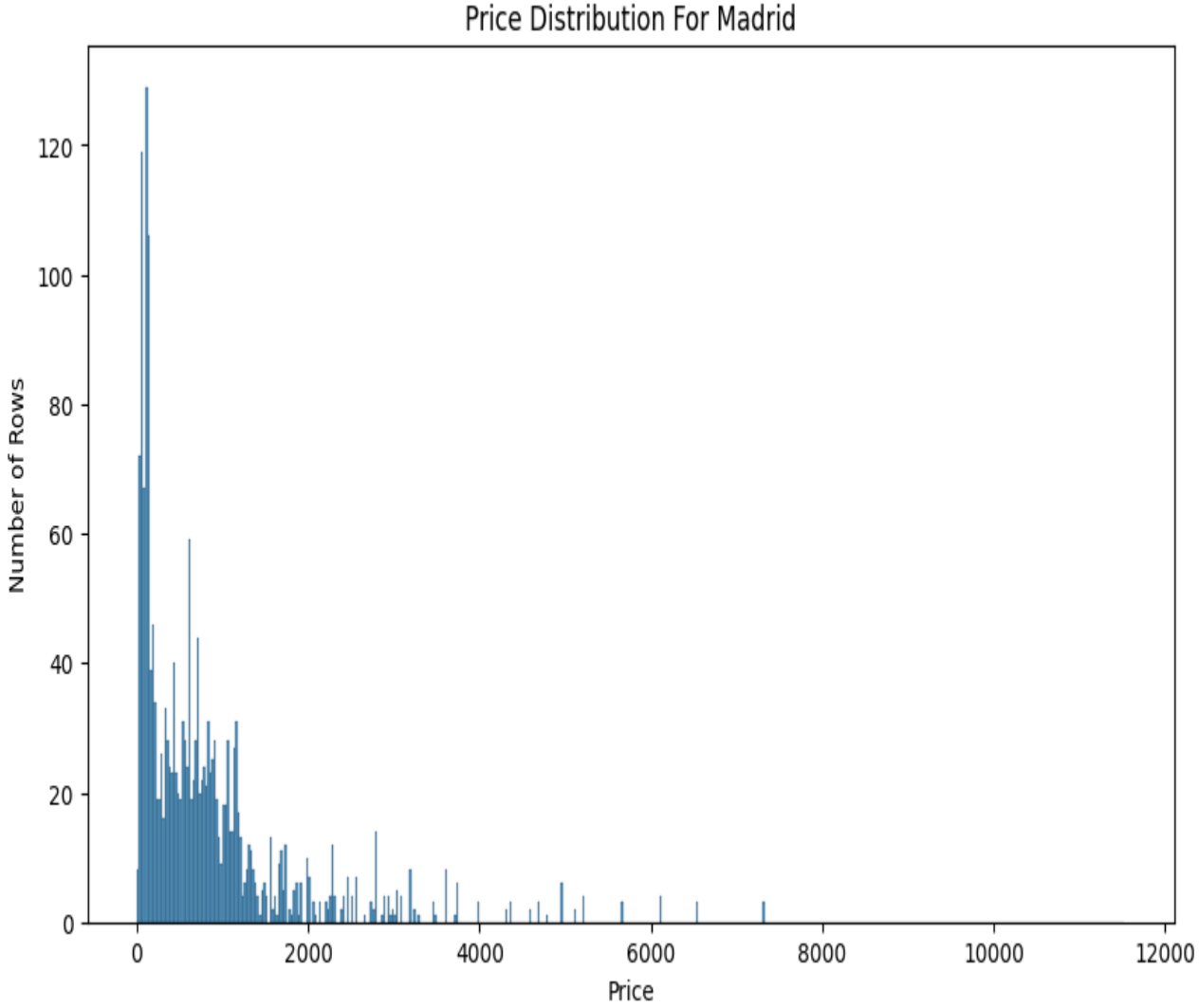




Madrid

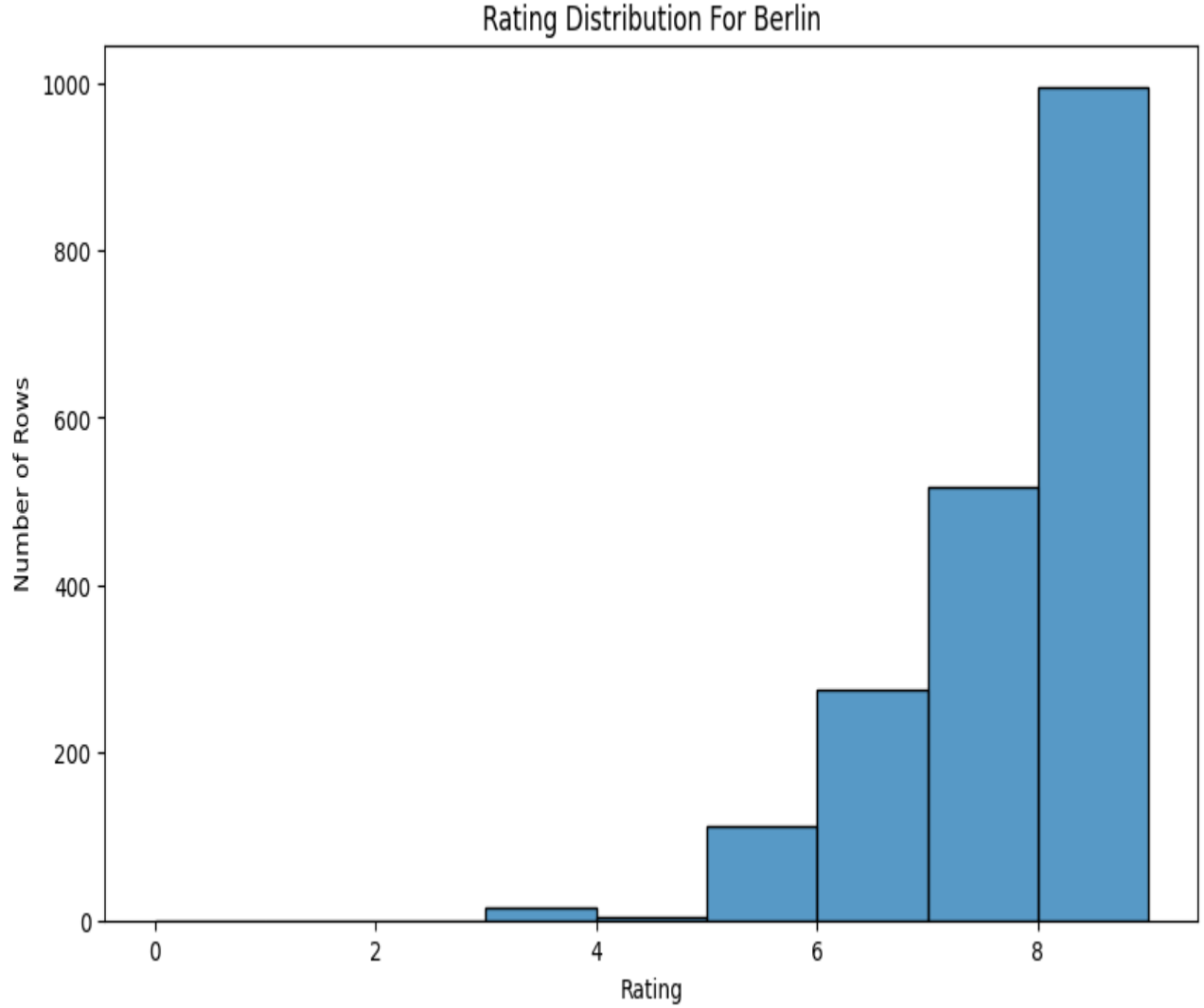
Rating and Price distribution

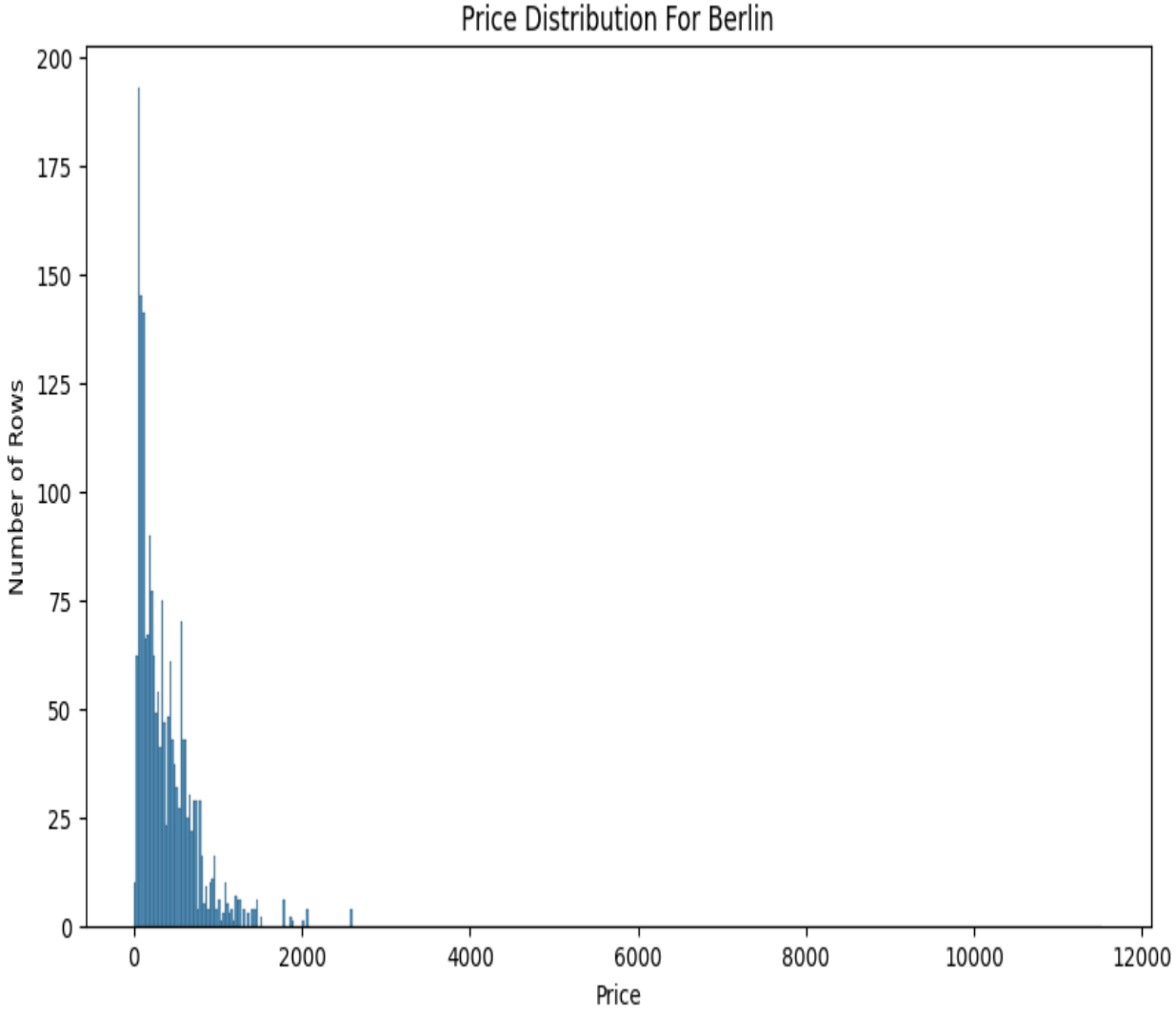




Berlin

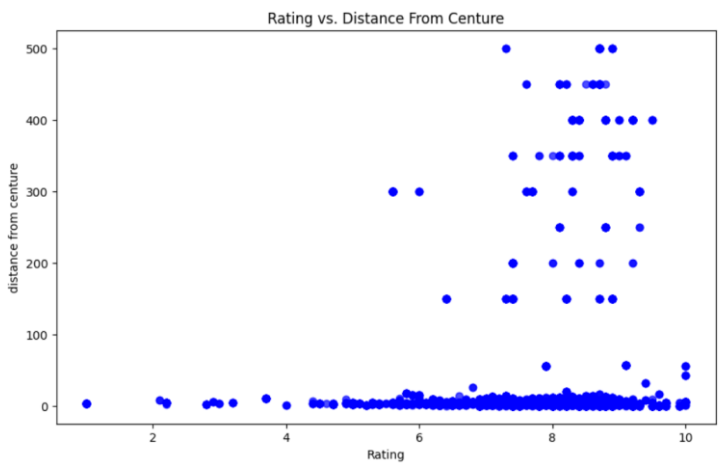
Rating and Price distribution

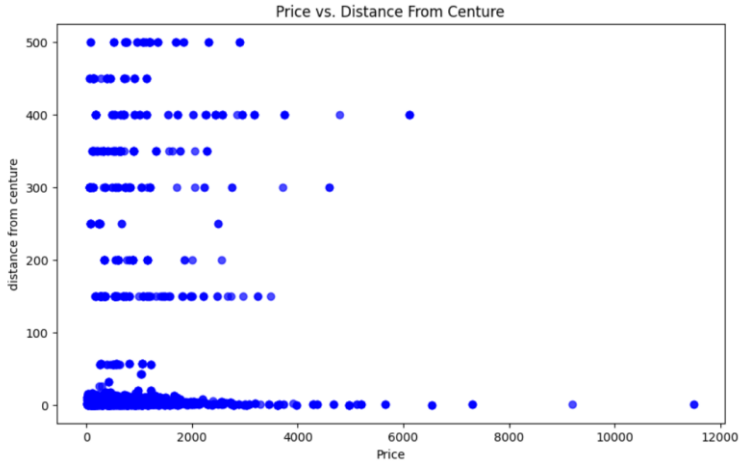


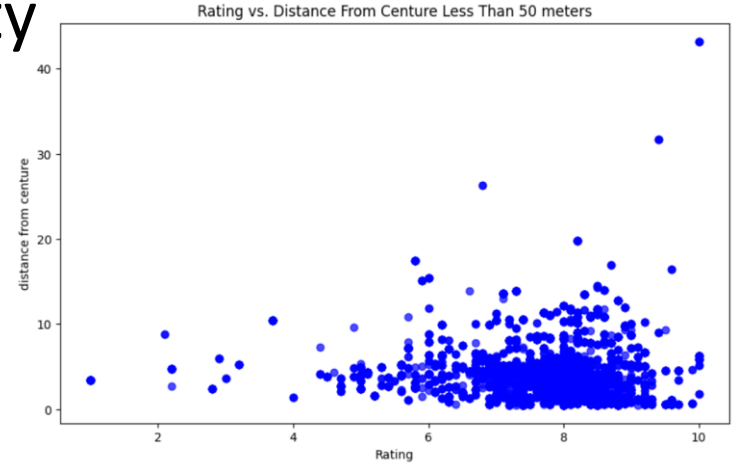


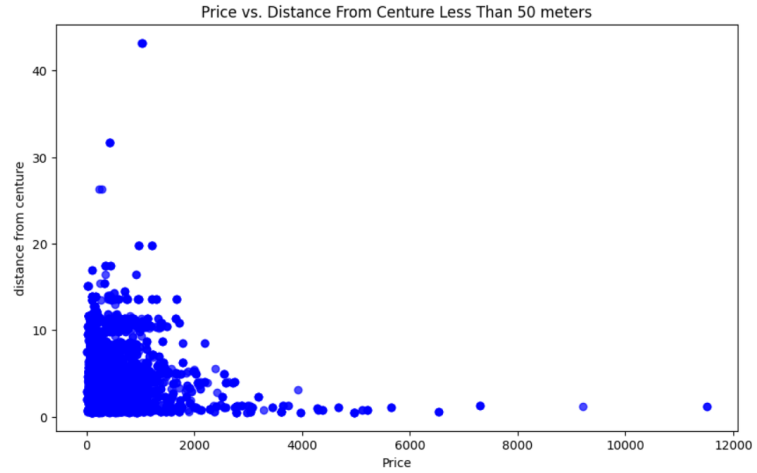
Visualization (Scatter Plot of Distance From Center)

- The best hotels are in center of cities mostly.
- Far hotels from center of city are cheaper.
- Don't worry about the poor quality of hotels by moving away from the city center
- So if you want cheaper hotel far places sound good idea.



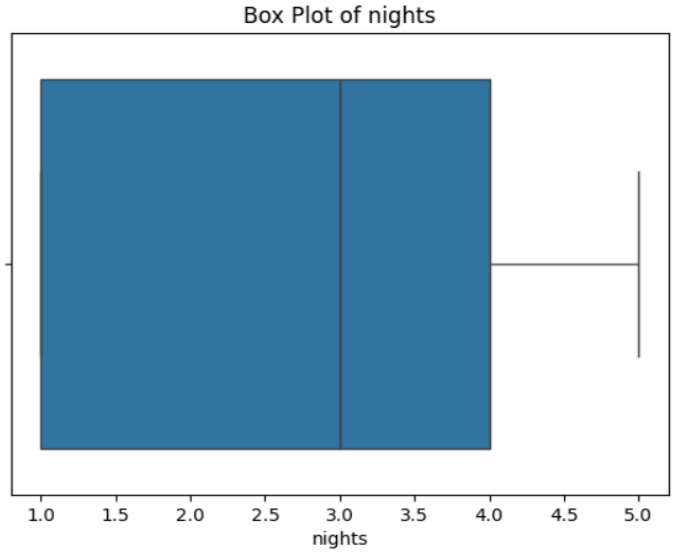


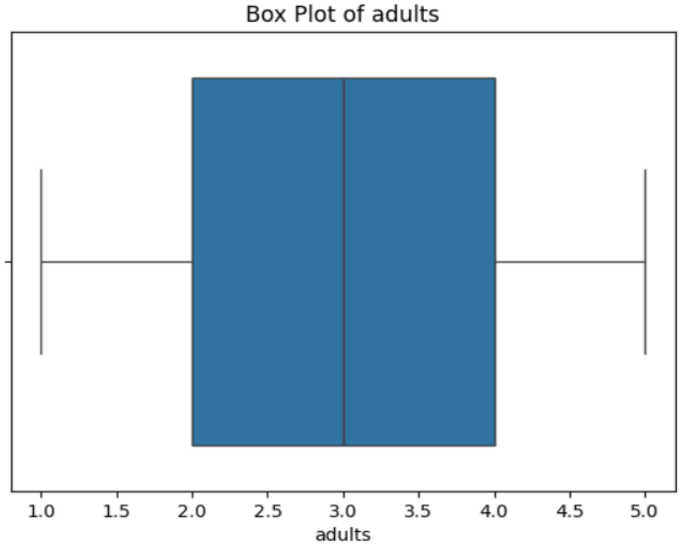


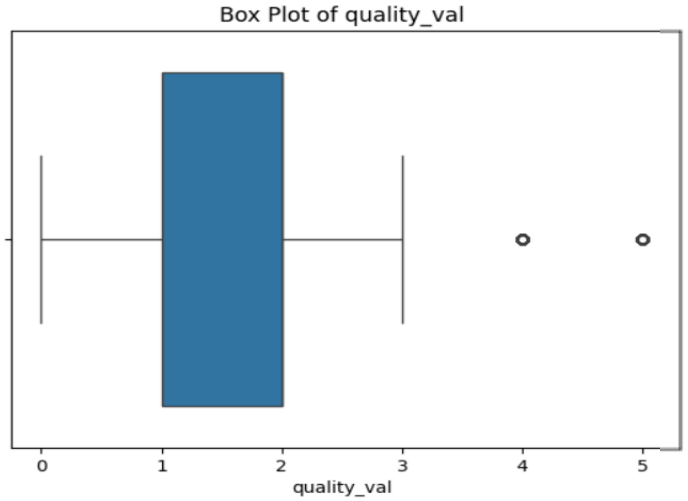


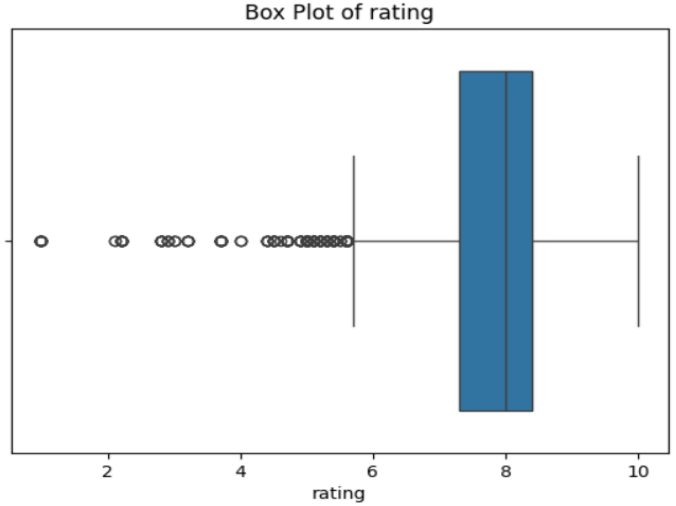
Visualization (Box Plots)

- The number of nights and adults are relatively stable with no outliers.
- The quality value has a few higher outliers, suggesting variability in the quality ratings provided.
- The ratings show a notable number of lower outliers, which could indicate occasional dissatisfaction among some guests despite the generally high median rating.
- There are numerous outliers below the rating of 6, showing that while most ratings are high, there are some significantly lower ratings present in the dataset.
- There are several outliers above the value of 4, indicating some higher quality values that are not common in the dataset.



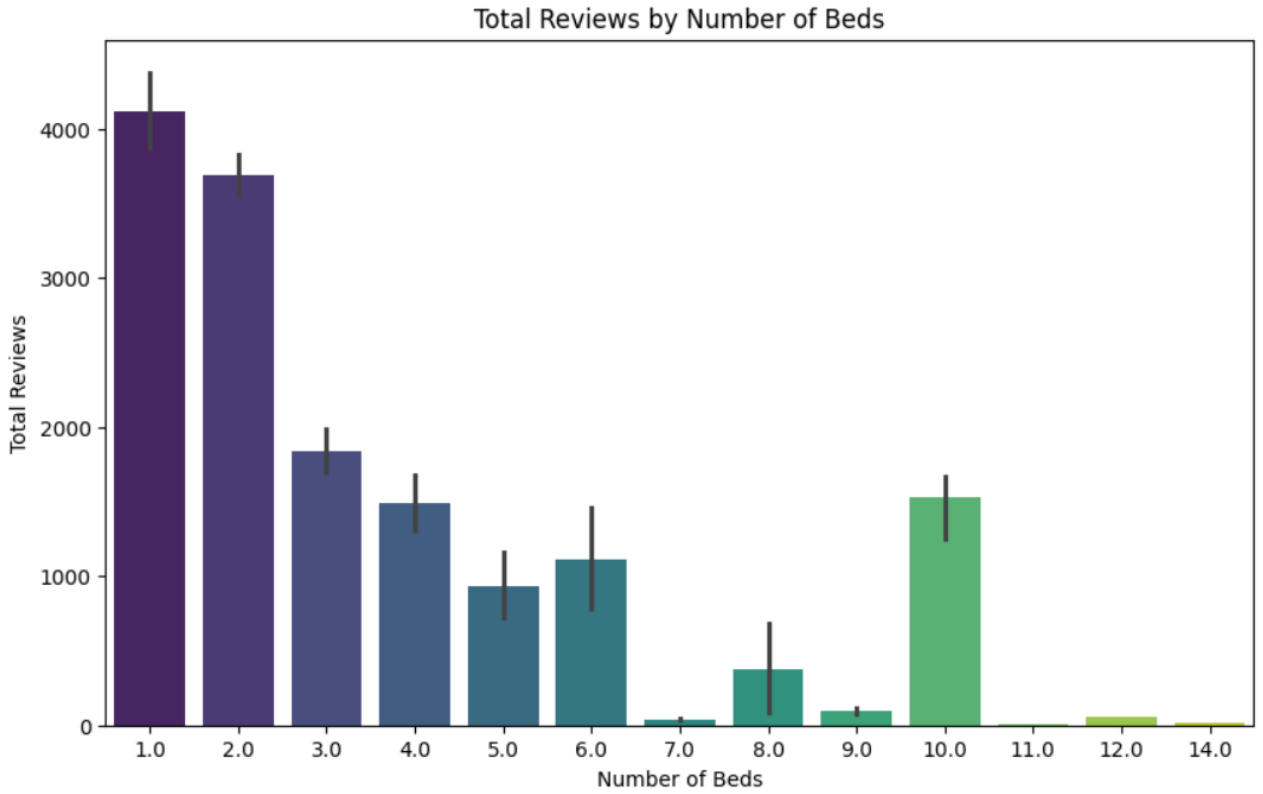






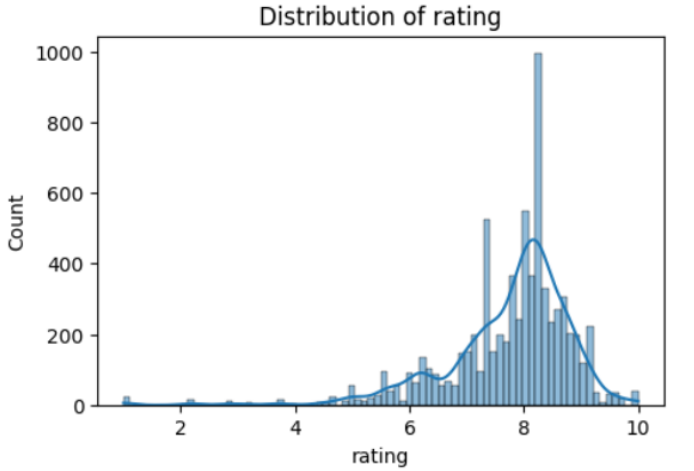
Visualization

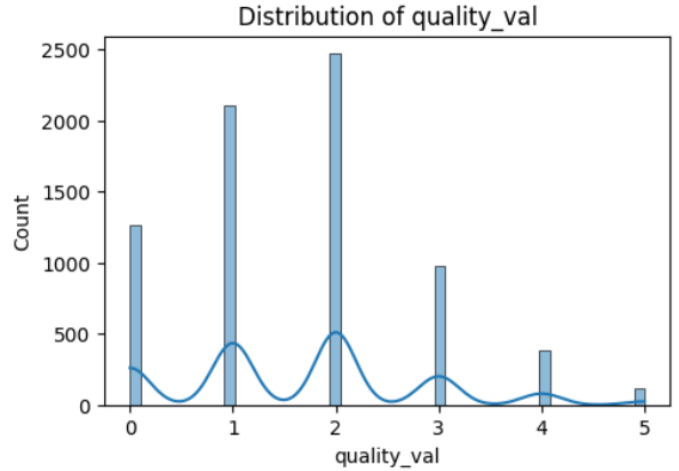
- Rooms with two beds have more reviews maybe because pepole in travel are family are group , so these reservations are more than ets.
- And maybe because reaction in groups or family are more than solo or couple and reactions are more.

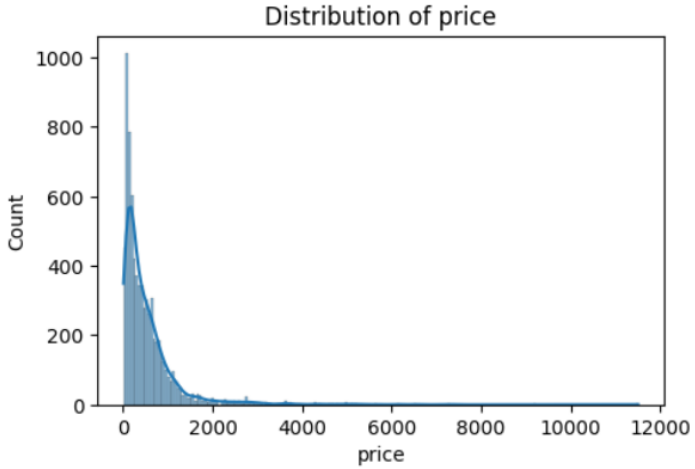


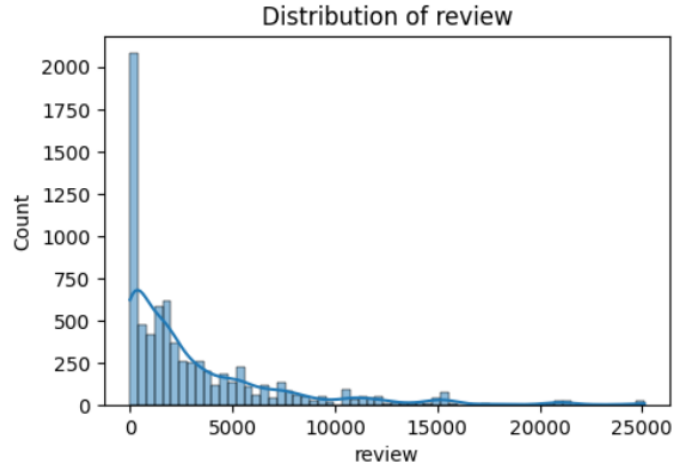
Visualization (Distributions)

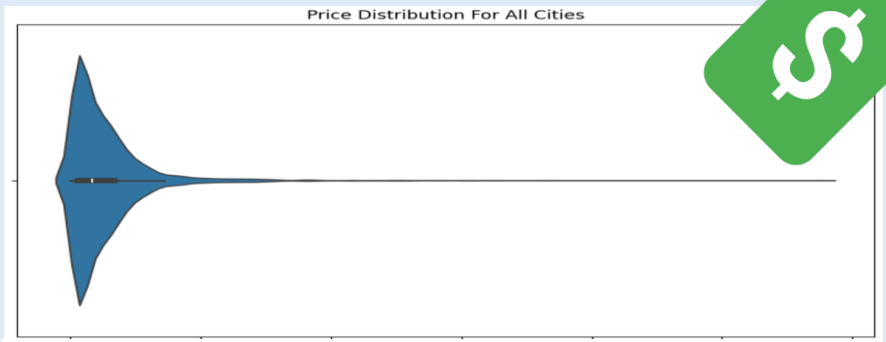
- The rating distribution shows a positive skew, indicating that most guests rate their experiences highly.
- The quality values are assigned in discrete categories, with most properties falling in the middle categories.
- Both price and review distributions are right-skewed, indicating that while most properties are priced reasonably and have a moderate number of reviews, there are a few properties that are significantly more expensive or have an unusually high number of reviews.
- There are some properties with an exceptionally high number of reviews, reaching up to 25000, but these are outliers.
- The price distribution is heavily right-skewed with a long tail. Most prices are concentrated below 2000 units, with a sharp drop-off as prices increase. A few data points show prices as high as 12000 units, but these are rare.

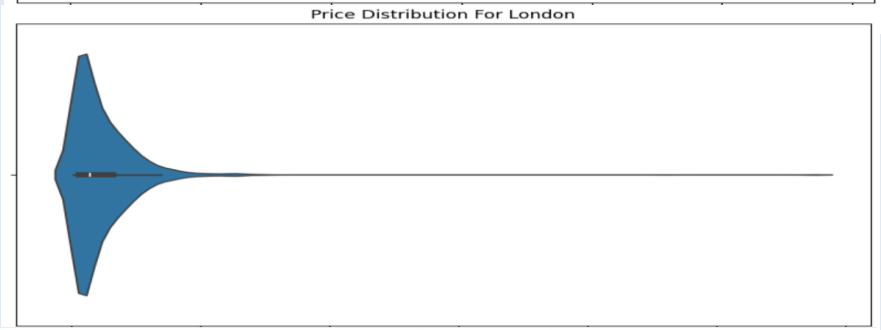


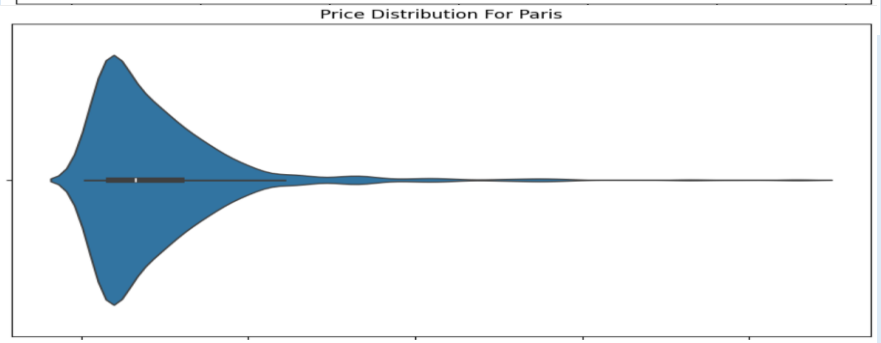


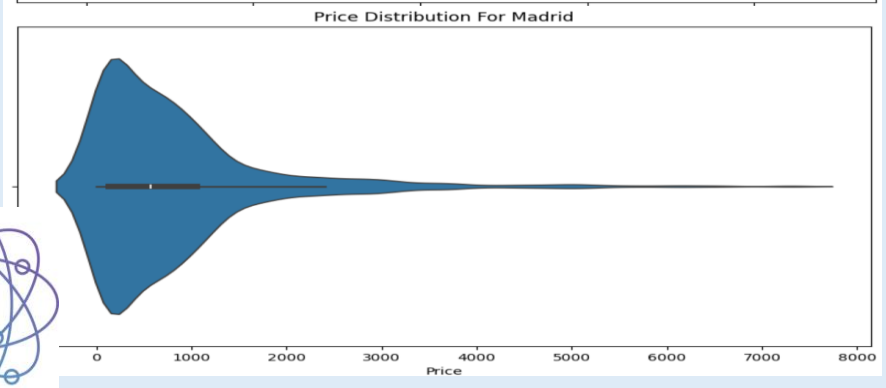


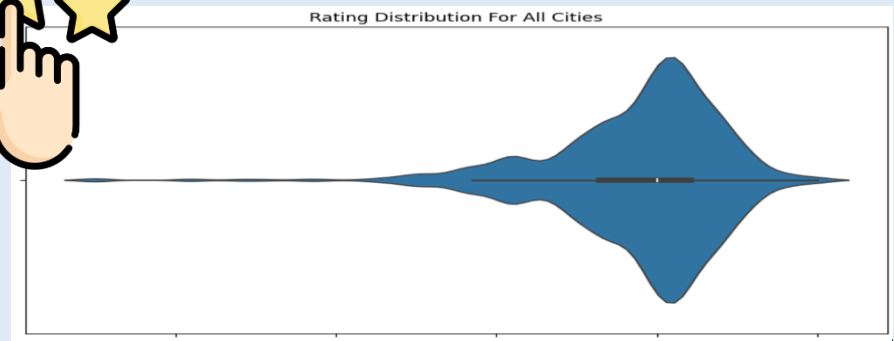


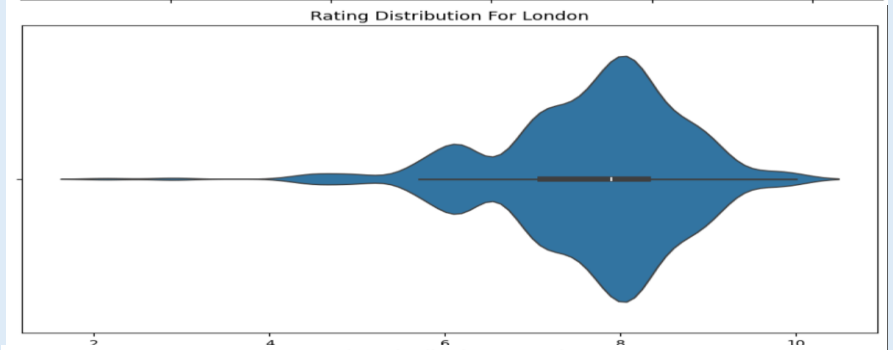

Price Distribution For All Cities

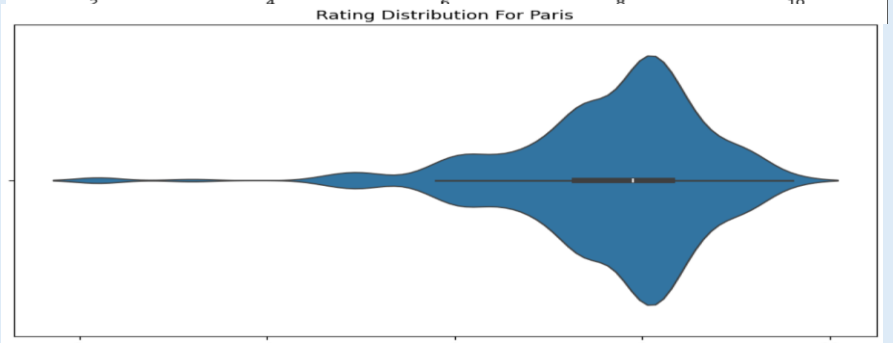

Price Distribution For London

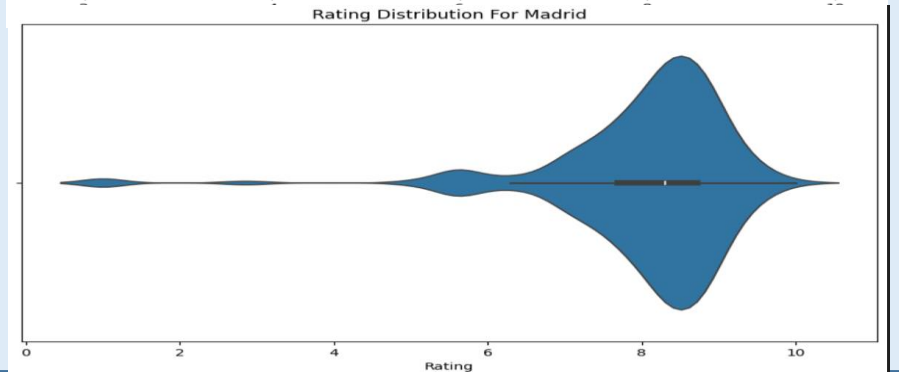

Price Distribution For Paris


Price Distribution For Madrid
0
1000
2000
3000
4000
5000
6000
7000
8000
Price


Rating Distribution For All Cities


Rating Distribution For London


Rating Distribution For Paris

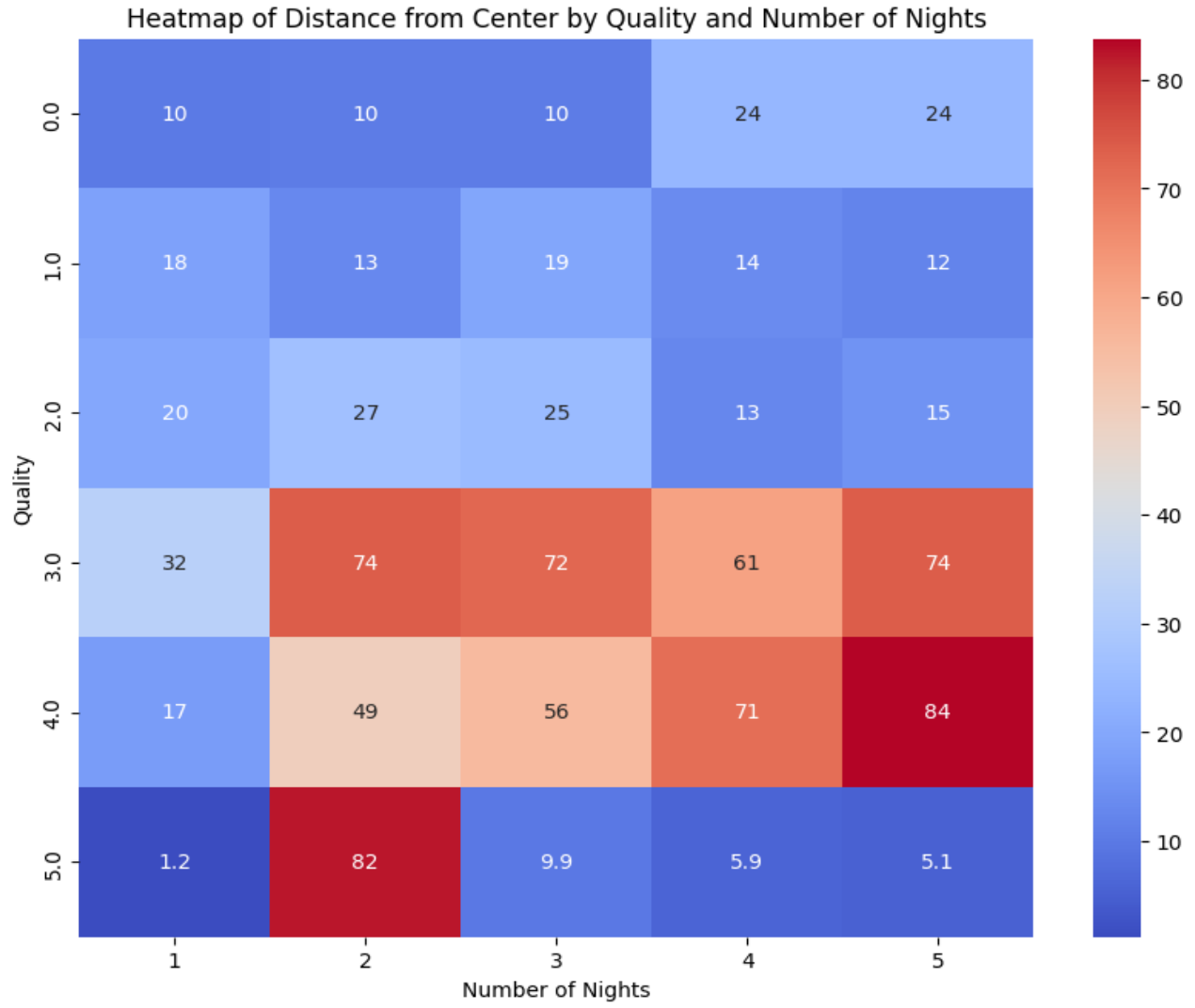

Rating Distribution For Madrid
0
2
4
6
8
10
Rating

Visualization (heatmap)

- Number of nights and quality heat map
- Red colors means higher price and vice versa

Visualization (heatmap)

- Number of nights and quality heat map
- Red colors means more distance from center of city and vice versa



OLS Result

- **Rating**: A highly significant positive effect (coef = 6.836e+05) indicating that higher ratings substantially increase the price.
- **Quality value:** Significant negative effect (coef = -2.537e+05), meaning that higher perceived quality decreases the price, which might suggest higher quality properties are priced lower to attract more bookings.
- **Nights:** Significant positive effect (coef = 817.50) suggesting that longer stays are associated with higher prices.
- **Adults:** Significant positive effect (coef = 533.70) indicating that the price increases with the number of adults.
- **Review:** Significant negative effect (coef = -165.95), meaning that better reviews lead to lower prices, which might indicate competitive pricing strategies.

- Various specific property names significantly influence the price, either positively or negatively. For example:
  - Positive Effects:
    - "134 Kilburn road apartment" (coef = 9.549e+04)
    - "250 City Road 2 Bollinder Place" (coef = 6.274e+04)
  - Negative Effects:
    - "Suite Ober - Superbe Appartement à Paris" (coef = -1.75e+05)
    - "2 Bed 2 Bath Close to Big Ben" (coef = -4.607e+05)

Tnx 4 your attention

Have a nice trip!